Simulation of single cell RNA-seq count data

Giacomo Baruzzo, PhD

PostDoc @ Department of Information Engineering

University of Padova

Outline

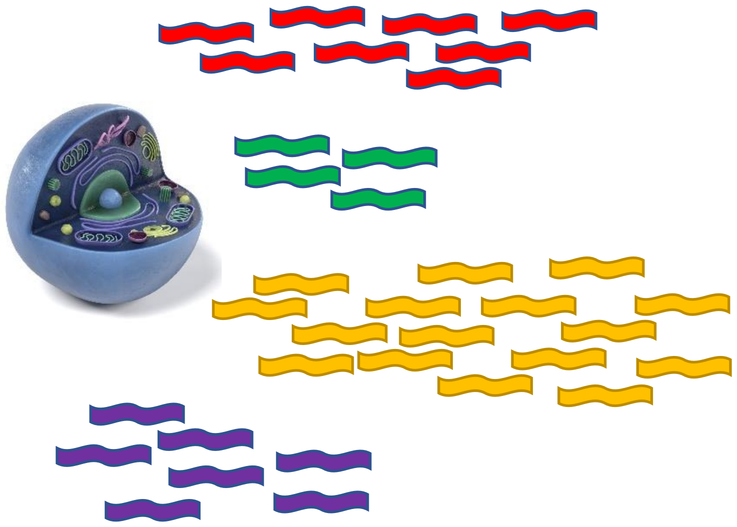
- Gene expression level
  - What it is and why it is important
  - Gene expression data
  - How to measure it: single cell RNA-seq
- Single cell RNA-seq
  - Experimental and bioinformatics workflow
  - Count data
  - Biases in measuring
- Simulating scRNA-seq count data
  - Why simulating data?
  - SPARSim simulator
  - Example of simulation

Goals

- Understand important concepts
  - Gene expression level
  - Biological variability
  - Technical variability (technical biases)
  - Data simulation

- Understand scRNA-seq count data
  - Know your data
  - “Play” with (simulated) scRNA-seq count data

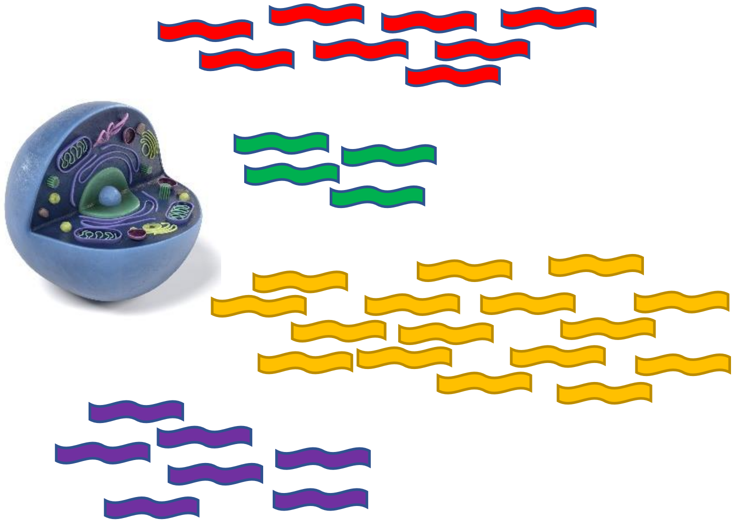
Gene expression level



Gene 1	8
Gene 2	4
Gene 3	0
Gene 4	16
…	…
Gene N	7

Gene expression level: amount of RNA molecules “produced” by each gene in a cell

Why studying gene expression level



Gene 1	8
Gene 2	4
Gene 3	0
Gene 4	16
…	…
Gene N	7

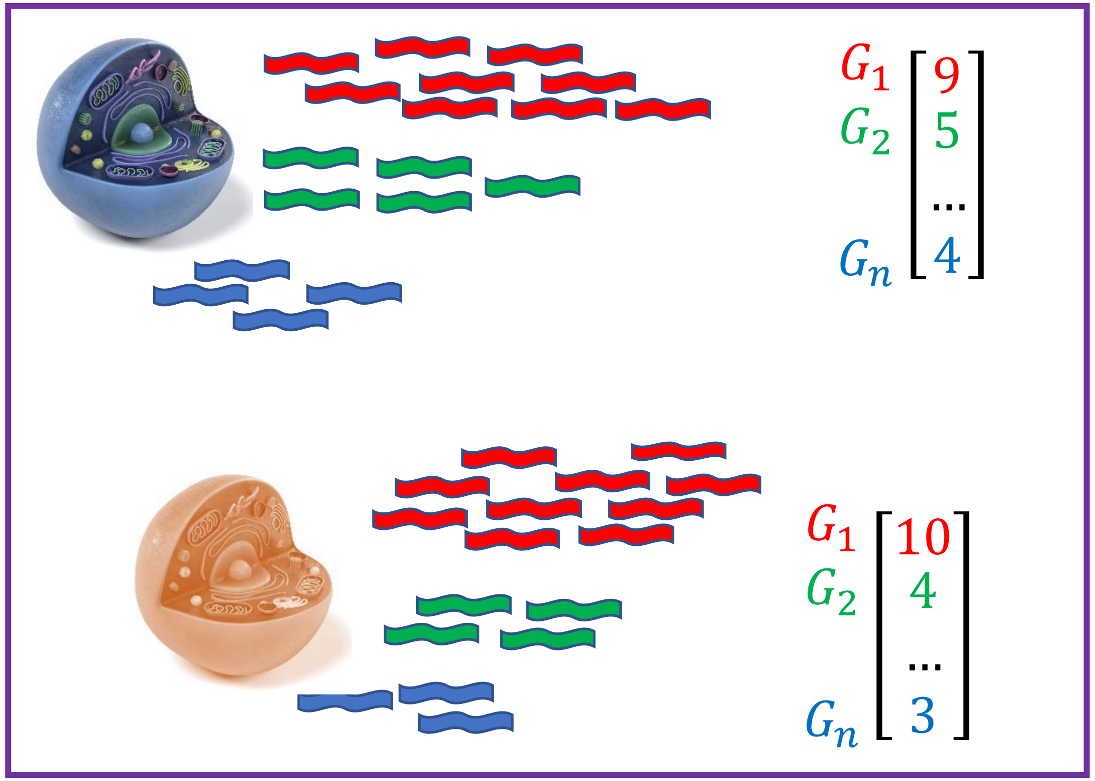
Low expressed gene

Turned-off gene

- Gene expression level describes the **state of a cell** and **what a cell is doing:**
  - **which genes are on/off**
  - **intensity of RNA production**

Why studying gene expression level

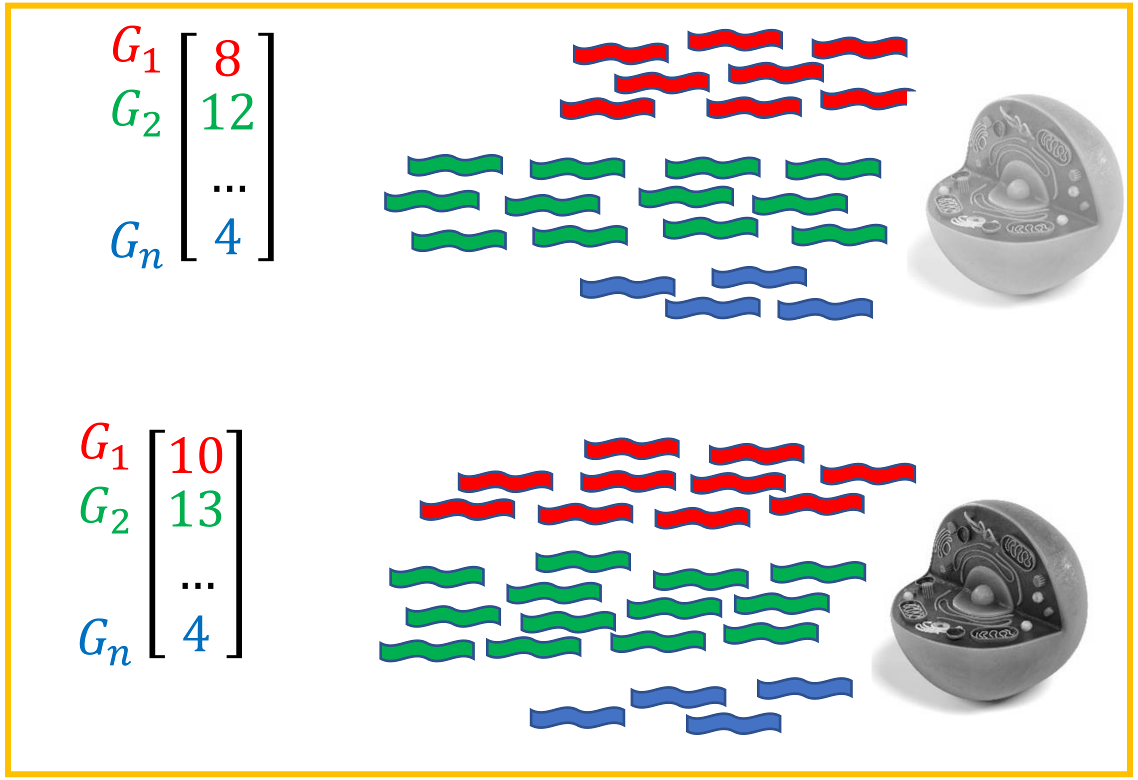
Experimental condition A (e.g. healthy cells)



$$\begin{array}{c} G_1 \\ G_2 \\ \\ G_n \end{array}\begin{bmatrix} 9 \\ 5 \\ \dots \\ 4 \end{bmatrix}$$

$$\begin{array}{c} G_1 \\ G_2 \\ \\ G_n \end{array}\begin{bmatrix} 10 \\ 4 \\ \dots \\ 3 \end{bmatrix}$$

Experimental condition B (e.g. cancer cells)

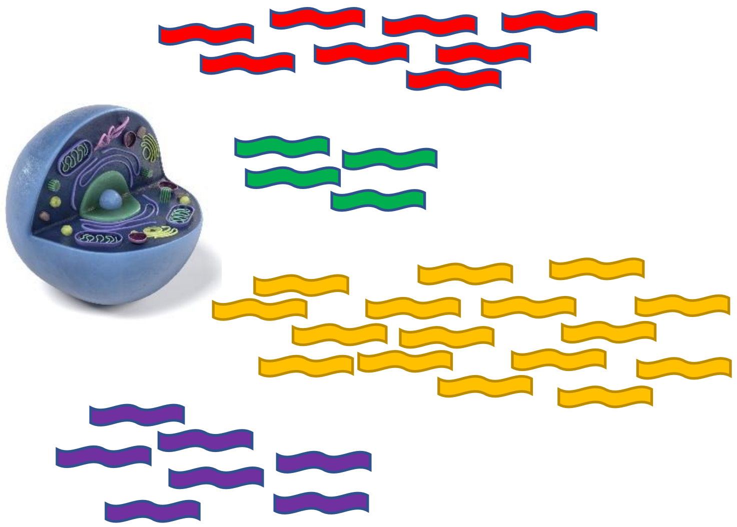


$$\begin{array}{c} G_1 \\ G_2 \\ \\ G_n \end{array}\begin{bmatrix} 8 \\ 12 \\ \dots \\ 4 \end{bmatrix}$$

$$\begin{array}{c} G_1 \\ G_2 \\ \\ G_n \end{array}\begin{bmatrix} 10 \\ 13 \\ \dots \\ 4 \end{bmatrix}$$

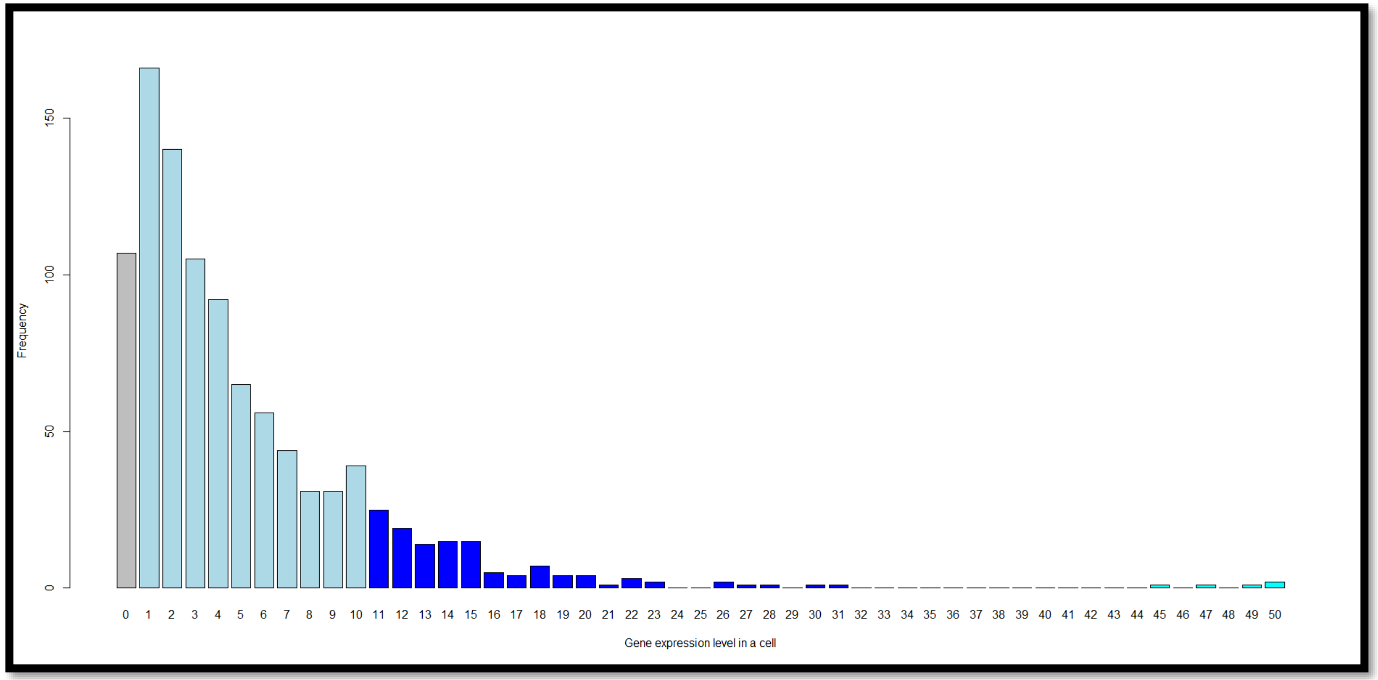
- Gene G_2 is **differentially expressed among healthy cells and cancer cells**…
- … then gene G_2 has a role in cancer

More about gene expression level



Gene 1	8
Gene 2	4
Gene 3	0
Gene 4	16
…	…
Gene N	7

- In a cell:
  - **Some genes** with a **null expression level**
  - **Many genes** with a **low expression level**
  - **Some genes** with a **medium expression level**
  - **Very few genes** with a **high expression level**



Gene expression level of many cells

Gene 1	10
Gene 2	3
Gene 3	0
Gene 4	75
Gene 5	47
…	…
Gene N	42

Gene 1	9
Gene 2	2
Gene 3	0
Gene 4	87
Gene 5	50
…	
Gene N	40

Gene 1	10
Gene 2	4
Gene 3	0
Gene 4	80
Gene 5	53
…	
Gene N	37

Gene 1	7
Gene 2	0
Gene 3	0
Gene 4	77
Gene 5	45
…	
Gene N	5

Matrix of gene expression level

Gene 1	10	9	10	7
Gene 2	3	2	4	0
Gene 3	0	0	0	0
Gene 4	75	87	80	77
Gene 5	47	50	53	45
…	…			
Gene N	42	40	37	5

- What a **group of cells** is doing

Gene expression level of many cells

Biological variability:

- **Cells in the same conditions** has **different expression levels**
- The amount of **variability** (*dispersion*) and the "**shape**" of the expression levels **depends on the intensity**

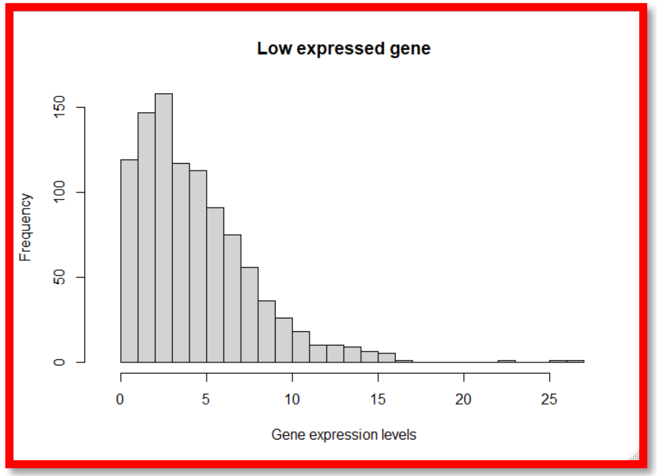
Gene 1	3	2	4	0	1	…	3
Gene 2	10	9	10	7	11		8
Gene 3	55	57	50	57	51		58
Gene 4	0	0	0	0	0		0
Gene 5	77	80	83	85	82		81
…	…					…	
Gene N	42	40	37	5	7		8

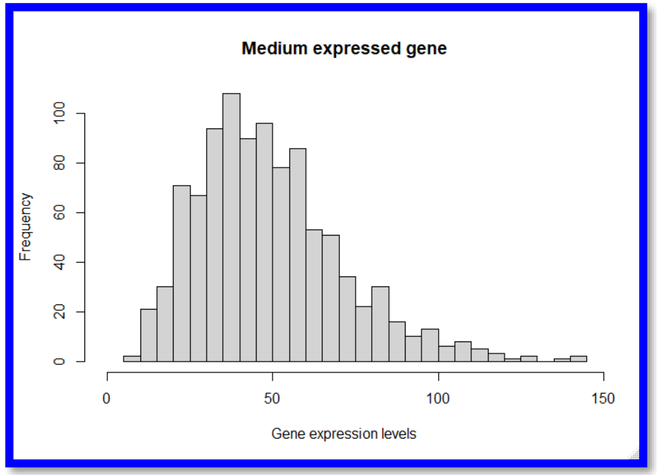
Gene expression level of many cells

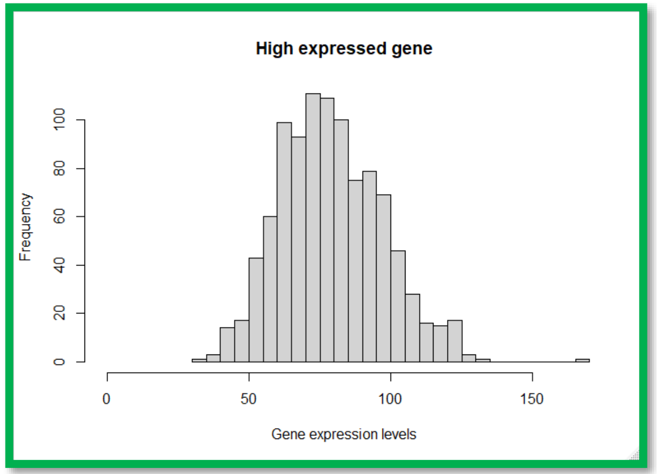
Biological variability:

- **Cells in the same conditions** has **different expression levels**
- The amount of **variability** (*dispersion*) and the “**shape**” of the expression levels **depends on the intensity**

Gene 1	3	2	4	0	1	…	3
Gene 2	10	9	10	7	11		8
Gene 3	55	57	50	57	51		58
Gene 4	0	0	0	0	0		0
Gene 5	77	80	83	85	82		81
…	…					…	
Gene N	42	40	37	5	7		8







Gene expression level of many cells, across different conditions

In case of multiple conditions (e.g. different cell types)

- The great majority of genes has similar genes expression levels
- Some genes are ***differentially expressed*** between different conditions

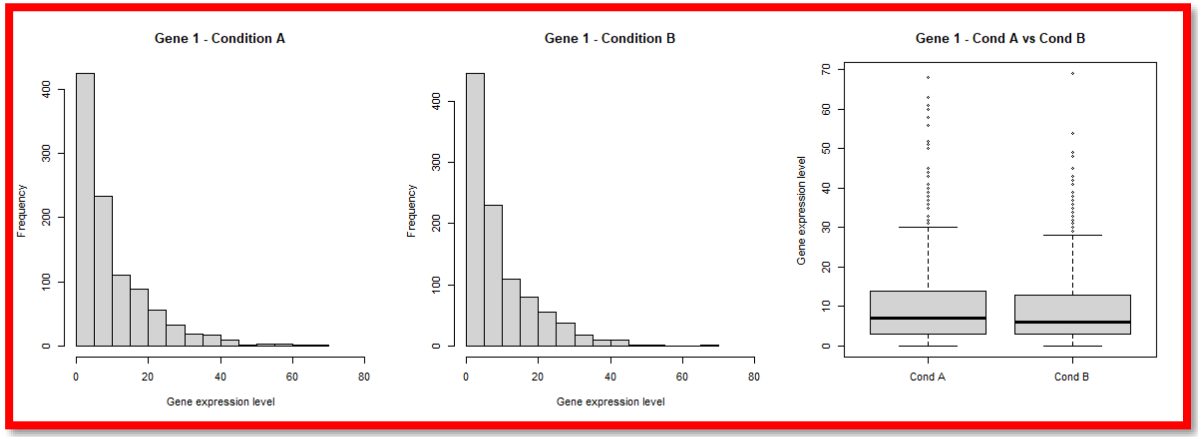
	Condition A				Condition B			
Gene 1	10	9	10	…	7	11	8	…
Gene 2	3	2	4	…	0	4	3	…
Gene 3	0	0	0	…	0	0	0	…
Gene 4	75	87	80	…	77	81	78	…
Gene 5	47	50	53	…	0	0	0	…
…	…			…				
Gene N	9	5	7	…	32	30	27	…

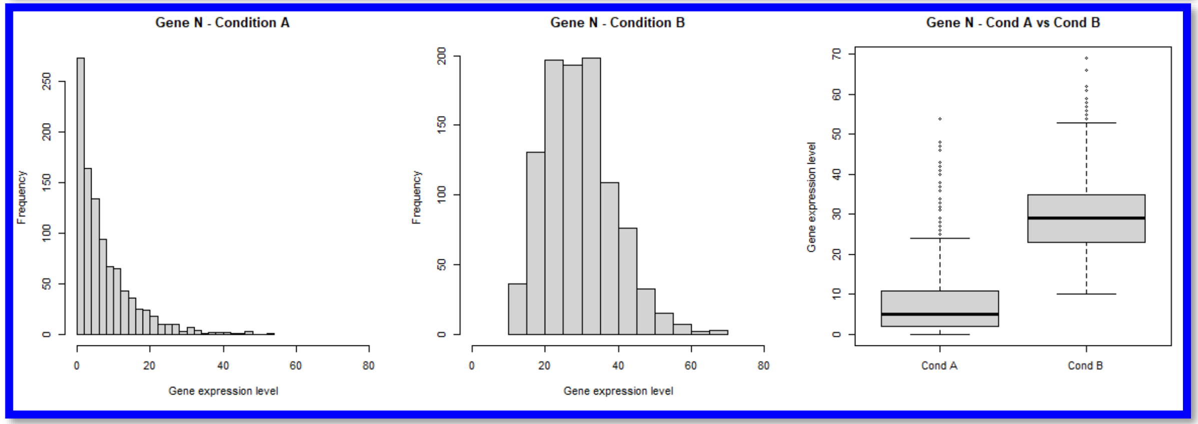
Gene expression level of many cells, across different conditions

In case of multiple conditions (e.g. different cell types)

- The great majority of genes has similar genes expression levels
- Some genes are ***differentially expressed*** between different conditions

	Condition A				Condition B			
Gene 1	10	9	10	…	7	11	8	…
Gene 2	3	2	4	…	0	4	3	…
Gene 3	0	0	0	…	0	0	0	…
Gene 4	75	87	80	…	77	81	78	…
Gene 5	47	50	53	…	0	0	0	…
…	…			…				
Gene N	9	5	7	…	32	30	27	…

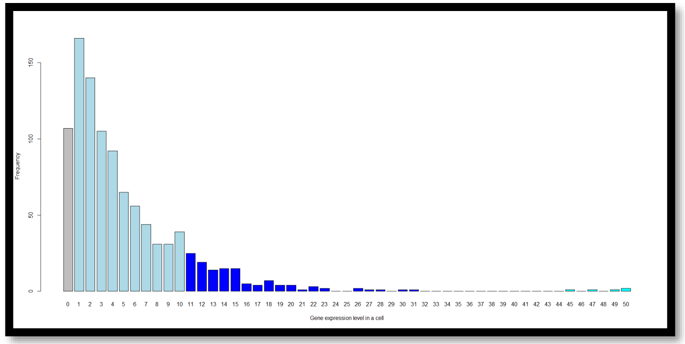




Gene expression level – Summary

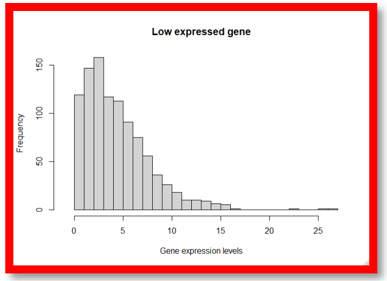
Each cell has:

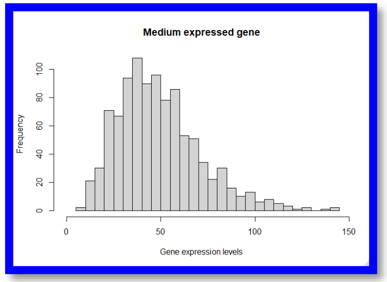
Skewed internal distribution of RNA molecules

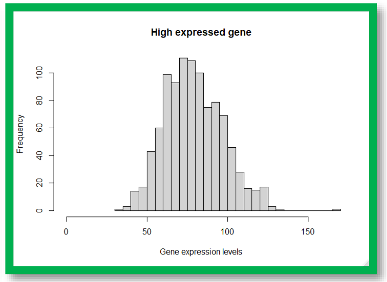


Cells in the same conditions:

Biological variability (dispersion, shape, …)



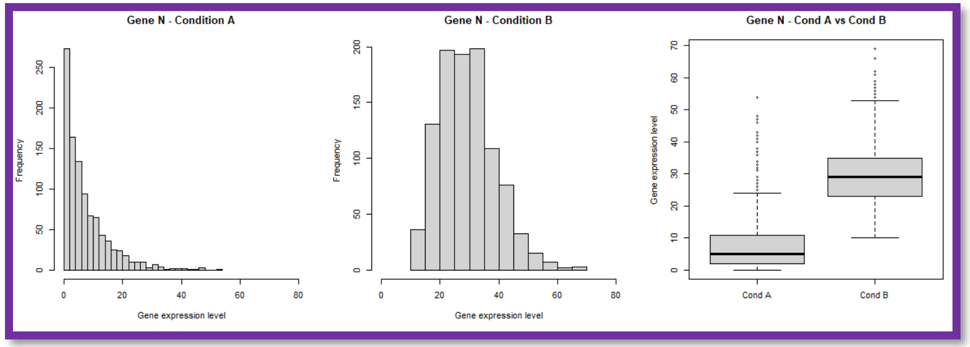




Cells in different conditions:

Some genes are differentially expressed

The great majority of genes are similar



Gene expression level

- Gene expression level is a very important information
  - It tells us what a cell is doing
  - It tells us what groups of cells are doing

- How can we measure the gene expression level?

- There are many ways to do it…
- … one of the most used and powerful one is **single cell RNA sequencing (scRNA-seq)**

Single cell RNA sequencing

Sample preparation

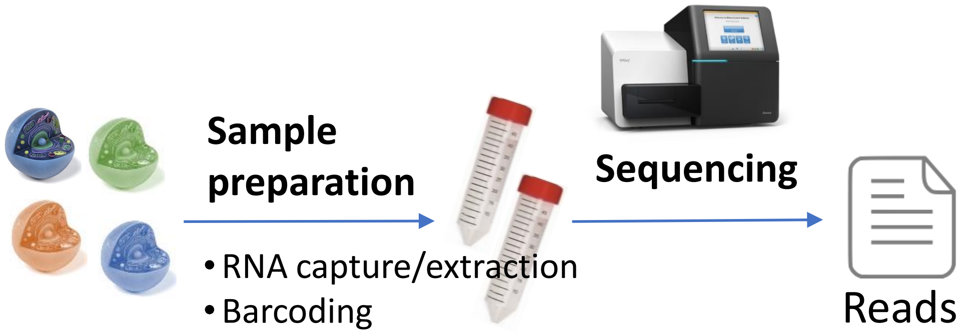
- RNA capture/extraction
- Barcoding
- Amplification
- …

Gene expression levels (X)

Gene 1	10	9	10	7
Gene 2	3	2	4	0
Gene 3	0	0	0	0
Gene 4	75	87	80	77
Gene 5	47	50	53	45
…	…			
Gene N	42	40	37	5

Unknown

Single cell RNA sequencing



Sample preparation

- RNA capture/extraction
- Barcoding
- Amplification
- …

Sequencing

Reads

Gene expression levels (X)

Gene 1	10	9	10	7
Gene 2	3	2	4	0
Gene 3	0	0	0	0
Gene 4	75	87	80	77
Gene 5	47	50	53	45
…	…			
Gene N	42	40	37	5

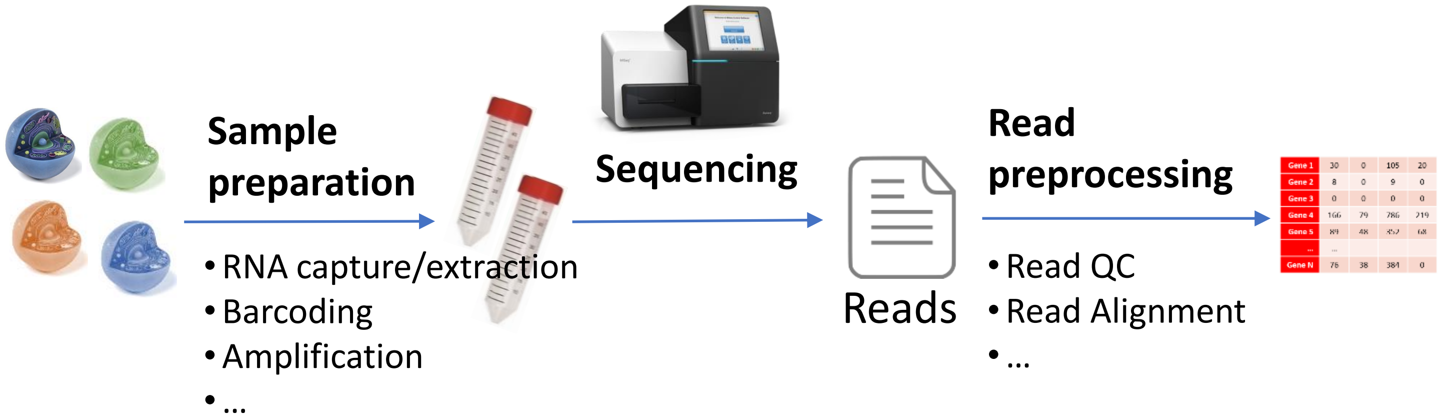
Unknown

Reads

```
>seq1_RTW_read1
ATCGACGTACGATGCACGCATGACG
>seq1_RTW_read2
ACGATGCATTGCATCGACTCGAATG
>seq1_RTW_read3
TTGCTAGTGTACCTGATGCATTGCA
>seq1_RTW_read4
CGACTCGAATACGATGCATTGCATG
>seq1_RTW_read5
GTACCTGATTGCTAGTTGCATTGCA
…
```

- Each string represents (a piece) of an RNA molecules

Single cell RNA sequencing



Sample preparation

- RNA capture/extraction
- Barcoding
- Amplification
- …

Sequencing

Reads

Read preprocessing

- Read QC
- Read Alignment
- …

Gene expression levels (X)

Gene 1	10	9	10	7
Gene 2	3	2	4	0
Gene 3	0	0	0	0
Gene 4	75	87	80	77
Gene 5	47	50	53	45
…	…			
Gene N	42	40	37	5

Unknown

Reads

```
>seq1_RTW_read1
ATCGACGTACGATGCACGCATGACG
>seq1_RTW_read2
ACGATGCATTGCATCGACTCGAATG
>seq1_RTW_read3
TTGCTAGTGTACCTGATGCATTGCA
>seq1_RTW_read4
CGACTCGAATACGATGCATTGCATG
>seq1_RTW_read5
GTACCTGATTGCTAGTTGCATTGCA
…
```

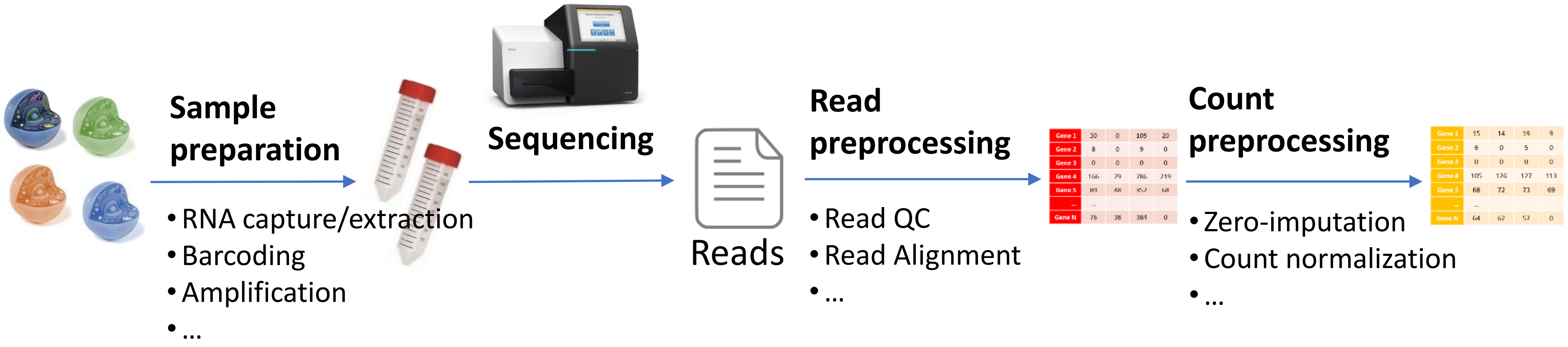
- Each string represents (a piece) of an RNA molecules

Raw count table (Y)

Gene 1	30	0	105	20
Gene 2	8	0	9	0
Gene 3	0	0	0	0
Gene 4	166	79	786	219
Gene 5	89	48	352	68
…	…			
Gene N	76	38	384	0

- Rough estimation of X
- Affected by biases
- # reads from gene i in cell j

Single cell RNA sequencing



Sample preparation

- RNA capture/extraction
- Barcoding
- Amplification
- …

Sequencing

Reads

Read preprocessing

- Read QC
- Read Alignment
- …

Count preprocessing

- Zero-imputation
- Count normalization
- …

Gene expression levels (X)

Gene 1	10	9	10	7
Gene 2	3	2	4	0
Gene 3	0	0	0	0
Gene 4	75	87	80	77
Gene 5	47	50	53	45
…	…			
Gene N	42	40	37	5

Unknown

Reads

```
>seq1_RTW_read1
ATCGACGTACGATGCACGCATGACG
>seq1_RTW_read2
ACGATGCATTGCATCGACTCGAATG
>seq1_RTW_read3
TTGCTAGTGTACCTGATGCATTGCA
>seq1_RTW_read4
CGACTCGAATACGATGCATTGCATG
>seq1_RTW_read5
GTACCTGATTGCTAGTTGCATTGCA
…
```

- Each string represents (a piece) of an RNA molecules

Raw count table (Y)

Gene 1	30	0	105	20
Gene 2	8	0	9	0
Gene 3	0	0	0	0
Gene 4	166	79	786	219
Gene 5	89	48	352	68
…	…			
Gene N	76	38	384	0

- Rough estimation of X
- Affected by biases
- # reads from gene i in cell j

Preprocessed count table ($\widetilde{Y}$)

Gene 1	15	14	19	9
Gene 2	6	0	5	0
Gene 3	0	0	0	0
Gene 4	105	126	127	113
Gene 5	68	72	73	69
…	…			
Gene N	64	62	57	0

- Better estimation of X
- Biases are mitigated

Single cell RNA sequencing

Sample preparation

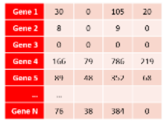
- RNA capture/extraction
- Barcoding
- Amplification
- …

Sequencing

Reads

Read preprocessing

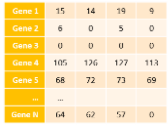
- Read QC
- Read Alignment
- …



Gene 1	30	0	105	20
Gene 2	8	0	9	0
Gene 3	0	0	0	0
Gene 4	166	79	786	219
Gene 5	89	48	352	68
…	…			
Gene N	76	38	384	0

Count preprocessing

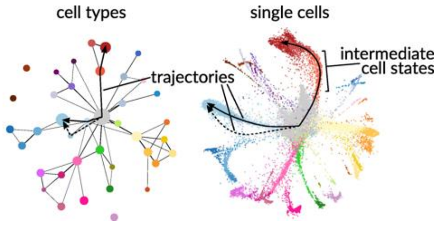
- Zero-imputation
- Count normalization
- …

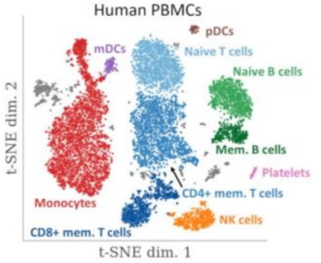


Gene 1	15	14	19	9
Gene 2	6	0	5	0
Gene 3	0	0	0	0
Gene 4	105	126	127	113
Gene 5	68	72	73	69
…	…			
Gene N	64	62	57	0

Analysis

- DE analysis
- Cells clustering
- ….





Gene expression levels (X)

Gene 1	10	9	10	7
Gene 2	3	2	4	0
Gene 3	0	0	0	0
Gene 4	75	87	80	77
Gene 5	47	50	53	45
…	…			
Gene N	42	40	37	5

Unknown

Reads

```
>seq1_RTW_read1
ATCGACGTACGATGCACGCATGACG
>seq1_RTW_read2
ACGATGCATTGCATCGACTCGAATG
>seq1_RTW_read3
TTGCTAGTGTACCTGATGCATTGCA
>seq1_RTW_read4
CGACTCGAATACGATGCATTGCATG
>seq1_RTW_read5
GTACCTGATTGCTAGTTGCATTGCA
…
```

- Each string represents (a piece) of an RNA molecules

Raw count table (Y)

Gene 1	30	0	105	20
Gene 2	8	0	9	0
Gene 3	0	0	0	0
Gene 4	166	79	786	219
Gene 5	89	48	352	68
…	…			
Gene N	76	38	384	0

- Rough estimation of X
- Affected by biases
- # reads from gene i in cell j

Preprocessed count table ($\widetilde{Y}$)

Gene 1	15	14	19	9
Gene 2	6	0	5	0
Gene 3	0	0	0	0
Gene 4	105	126	127	113
Gene 5	68	72	73	69
…	…			
Gene N	64	62	57	0

- Better estimation of X
- Biases are mitigated

Raw count table

Y is a matrix of ***N*** rows (genes) and ***M*** columns (cells)
$Y_{i,j}$ = # reads/UMIs of gene ***i*** in cell ***j*** (not negative integers)

It is just a matrix of numbers: cell condition/type is unknown

***M* cells**

***N* genes**

Gene 1	10	9	10	…	7
Gene 2	3	2	4	…	0
Gene 3	0	0	0	…	0
Gene 4	0	87	80	…	77
Gene 5	4	0	0	…	5
…	…	…	…	…	….
Gene N	42	0	0	…	54

Some characteristics of the count matrix depends on the technology used

New (i.e. last) technology
- measures more cells, but it detect less genes
- measured genes show lower technical biases

	Old technology	**New technology**
Number of genes (N)	Up to 15K	Up to 5K
Number of cells (M)	10^2-10^3	10^3-10^5
Sparsity	40%-80%	85%-97%

Raw count tables are (very) bad approximation of true gene expression level
The experimental procedure introduce many ***(technical) biases***:
- Some data is lost (e.g. a gene is not measured)
- Some data is changed (e.g. the measured value is wrong)

Technical biases - RNA capture/extraction

- Current RNA capture/extraction protocols have a low efficiency
  - Old protocols: capture 20% of RNA molecules
  - Current protocols: > 20% of RNA molecules

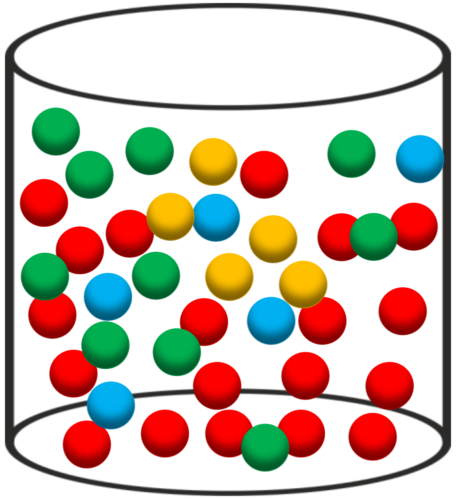
- The low efficiency introduces biases:
  - Not all the RNA molecule are captured
  - Low expressed genes may be not detected
  - All expression levels are “noisy”

Technical biases - RNA capture/extraction

- The low efficiency introduces biases:
  - Not all the RNA molecule are captured
  - Low expressed genes may be not detected
  - All expression levels are “noisy”

- An easy way to understand it
  - Urn (i.e. cell) with 40 balls (i.e. RNA molecule)
  - Balls of different colors (i.e. genes)
  - You can extract only 25% of balls (i.e. 10 balls)

Gene 1	20
Gene 2	10
Gene 3	5
Gene 4	5
Gene 5	0
Total	40



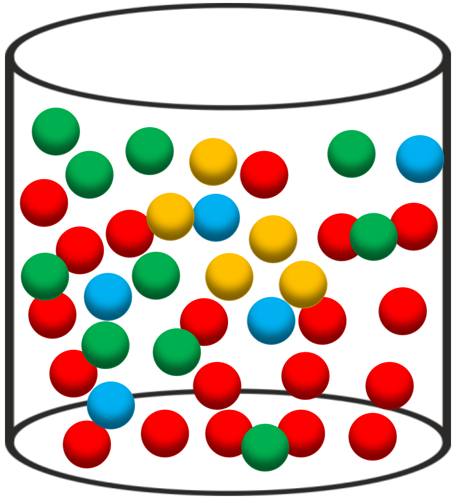
Technical biases - RNA capture/extraction

- The low efficiency introduces biases:
  - Not all the RNA molecule are captured
  - Low expressed genes may be not detected
  - All expression levels are “noisy”

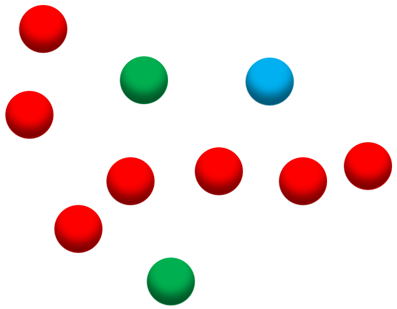
- An easy way to understand it
  - Urn (i.e. cell) with 40 balls (i.e. RNA molecule)
  - Balls of different colors (i.e. genes)
  - You can extract only 25% of balls (i.e. 10 balls)

- What happens?
  - Gene 4 is lost
  - Relative abundances inside the urn are wrong (e.g. Gene 2 is no more half of Gene 1)

Gene 1	20
Gene 2	10
Gene 3	5
Gene 4	5
Gene 5	0
Total	40



Gene 1	7
Gene 2	2
Gene 3	1
Gene 4	0
Gene 5	0
Total	10

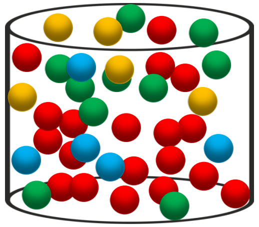


Technical biases - RNA capture/extraction

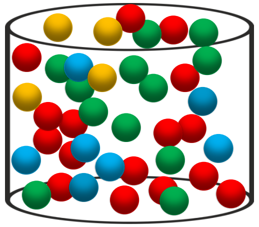
- The low efficiency introduces biases:
  - Not all the RNA molecule are captured
  - Low expressed genes may be not detected
  - All expression levels are “noisy”

- An easy way to understand it
  - **2 urns** (i.e. cell) with 40 balls (i.e. RNA molecule)
  - Balls of different colors (i.e. genes)
  - The **2 urns has similar amount of balls/colors**
  - You can extract only 25% of balls (i.e. 10 balls)

Gene 1	20
Gene 2	10
Gene 3	5
Gene 4	5
Gene 5	0
Total	40



Gene 1	17
Gene 2	12
Gene 3	7
Gene 4	4
Gene 5	0
Total	40



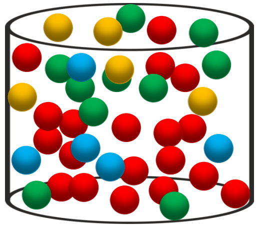
Technical biases - RNA capture/extraction

- The low efficiency introduces biases:
  - Not all the RNA molecule are captured
  - Low expressed genes may be not detected
  - All expression levels are “noisy”

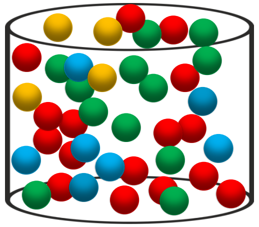
- An easy way to understand it
  - **2 urns** (i.e. cell) with 40 balls (i.e. RNA molecule)
  - Balls of different colors (i.e. genes)
  - The **2 urns has similar amount of balls/colors**
  - You can extract only 25% of balls (i.e. 10 balls)

- What happens?
  - Gene 4 is lost in urn 1, but not in urn 2
  - Relative abundances inside an urn are wrong (e.g. Gene 2 is no more half of Gene 1)
  - Relative abundances between urns are wrong (e.g. Gene 3 in urn 1 is not half of Gene 3 in urn 2)

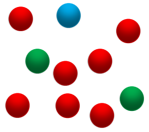
Gene 1	20
Gene 2	10
Gene 3	5
Gene 4	5
Gene 5	0
Total	40



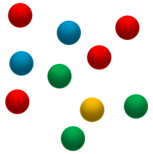
Gene 1	17
Gene 2	12
Gene 3	7
Gene 4	4
Gene 5	0
Total	40



Gene 1	7
Gene 2	2
Gene 3	1
Gene 4	0
Gene 5	0
Total	10



Gene 1	4
Gene 2	3
Gene 3	2
Gene 4	1
Gene 5	0
Total	10



Technical biases - RNA capture/extraction

RNA capture/extraction is a **competitive sampling with limited sampling size**:

- RNA molecules compete to be captured/extracted
- More abundant RNA molecules has more chance of beings captured
- Less abundant RNA molecules has more chance of being not captured

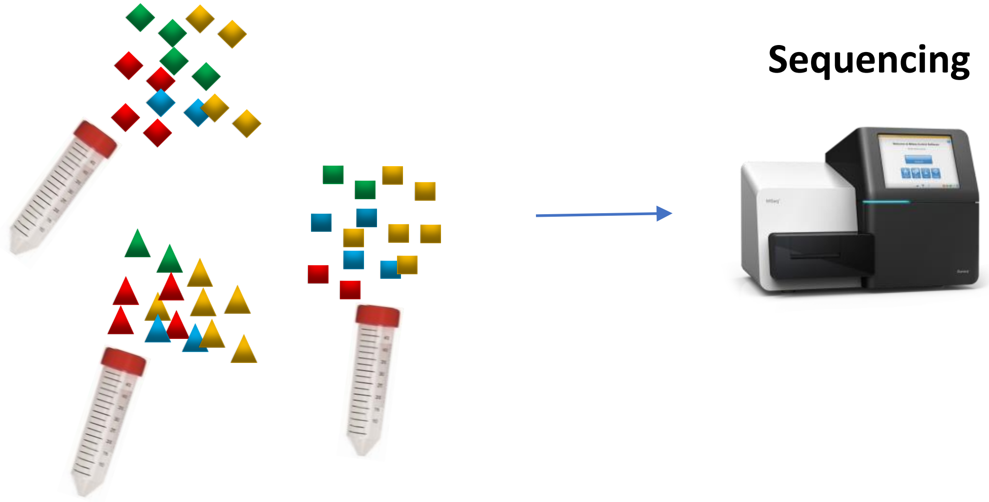
Effects:

- Some genes are lost (i.e. no RNA molecule is captured)
- The relative abundances inside a cell are changed
- The relative abundances between cells are changed

Technical biases - Sequencing process

- During the sequencing process, the RNA fragments provided as input of the sequencer are “read”
- Only a limited number of RNA fragments can be read
- RNA fragments “compete” to be read
- The limited number of reads and the completive sampling introduce biases
    - Low expressed RNA fragments may be not detected
    - Relative abundances are changed
    - Uneven number of reads per cells (aka sequencing depth or library size)

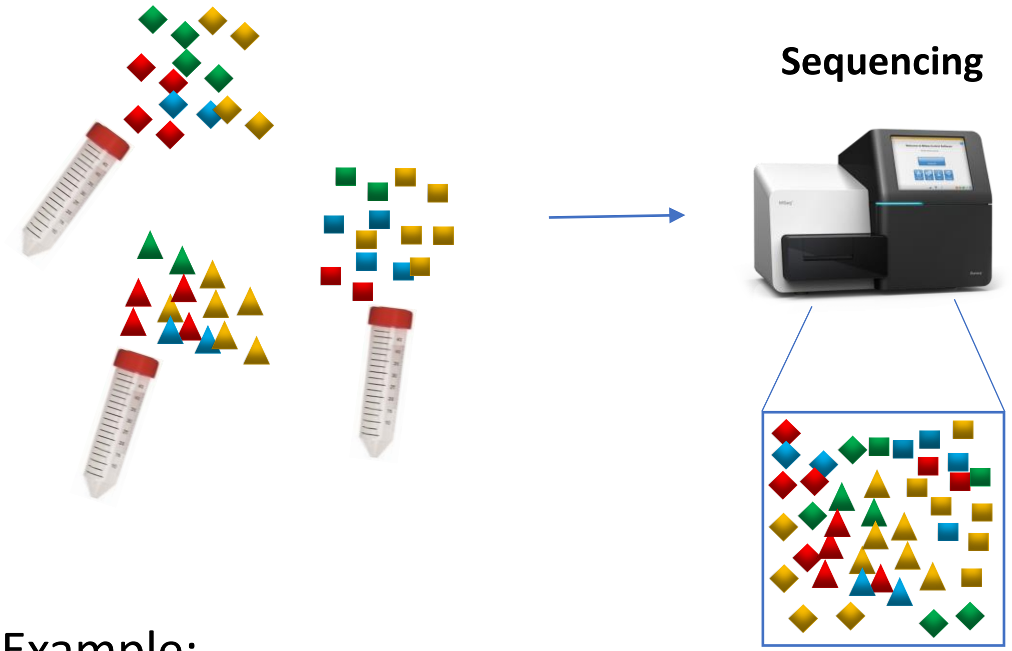
Technical biases - Sequencing process



Example:

- 3 cells (shapes)
- each cell has its own RNA fragments, coming from different genes (colors)

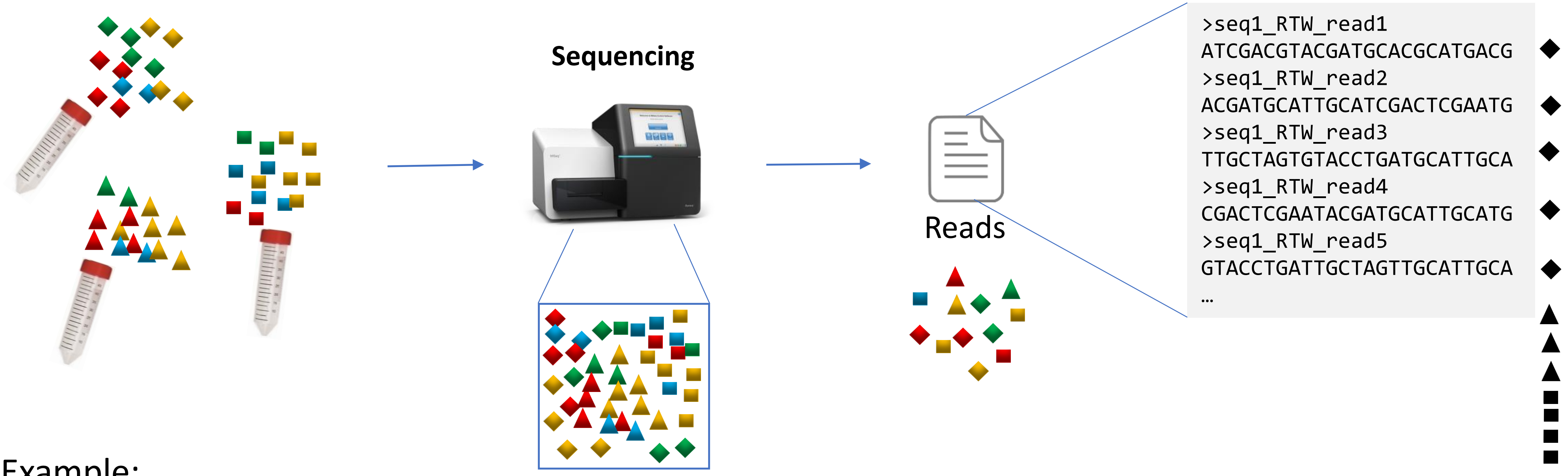
Technical biases - Sequencing process



Example:

- 3 cells (shapes)
- each cell has its own RNA fragments, coming from different genes (colors)

Technical biases - Sequencing process



```
>seq1_RTW_read1
ATCGACGTACGATGCACGCATGACG
>seq1_RTW_read2
ACGATGCATTGCATCGACTCGAATG
>seq1_RTW_read3
TTGCTAGTGTACCTGATGCATTGCA
>seq1_RTW_read4
CGACTCGAATACGATGCATTGCATG
>seq1_RTW_read5
GTACCTGATTGCTAGTTGCATTGCA
…
```

Example:

- 3 cells (shapes)
- each cell has its own RNA fragments, coming from different genes (colors)

Sequencing process is a competitive sampling: some fragments are lost, fragment abundances are changed, …

The amount of output RNA fragment (i.e. reads) of each cell (aka sequencing depth or library size) is different

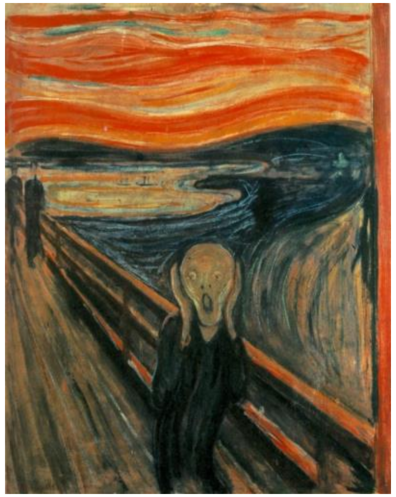
Technical biases - Summary

Experimental procedure introduces many biases:

- Some genes are not detected (i.e. technical zeros)
- The measured amount of RNA molecule is quite different from the true one

Raw count data are a bad approximation of true gene expression level

Gene expression level



Raw count

Technical biases - Summary

Raw counts are affected by technical biases

Technical variability: variability in the measured gene expression level (i.e. raw counts) due to technical biases.

Vs.

Biological variability: variability in the true gene expression level due to biological reasons.

Biological zeros vs technical zeros

Raw count matrix need pre-processing

Count pre-processing

- Set of methods, tools and software that remove/mitigate the technical biases
- Input: raw count matrix
- Output: pre-processed count matrix

- The are many scRNA-seq count pre-processing methods, each one remove/mitigate one or more biases

- Examples of scRNA-seq count pre-processing methods
    - Normalization
    - Zero-imputation
    - Batch effect removal
    - …

Pre-processed count matrix

Ideally, pre-processing should remove all the biases and the pre-processed count table should be proportional to the true (unknown) gene expression level

$$\boldsymbol{\widetilde{Y} = \alpha X}$$

In practice, pre-processing can **only mitigate the technical biases**…

… so a **good pre-processing** produces a pre-processed count table that is still an **approximation of the true (unknown) gene expression level**

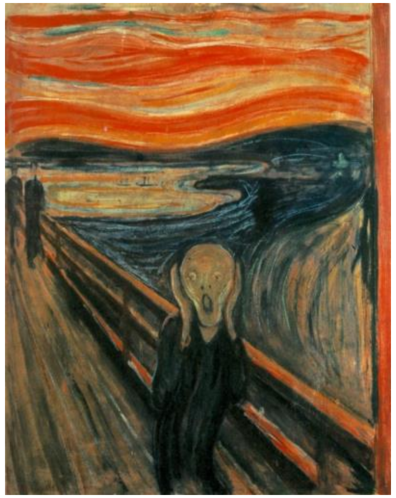
… but it is a better approximation compared with raw count table

Pre-processed count matrix

Single cell RNA-seq bioinformatics and data analysis:

- it is impossible to measure exactly the gene expression level…
- …do the best you can (i.e. do a good pre-processing)

Gene expression level

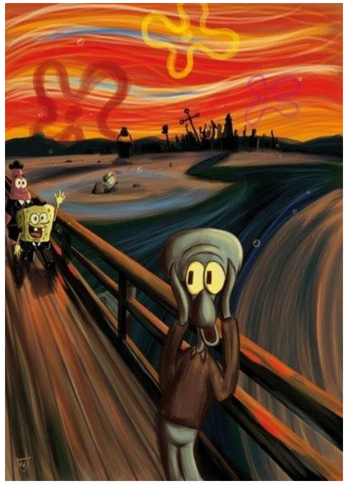


scRNA-seq experiments

Raw count

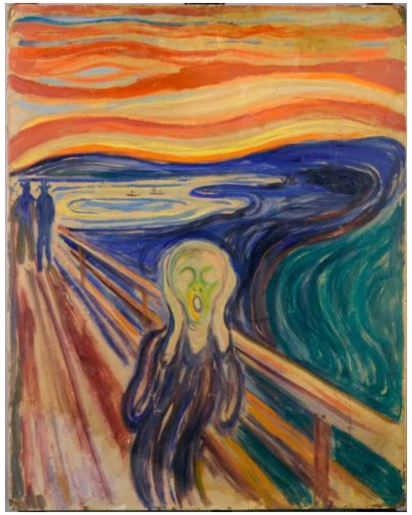
Bad pre-processing

Pre-processed count



Good pre-processing

Pre-processed count



Analysis

- Set of methods, tools and software that extract (biological) information from (pre-processed) data
- Input: pre-processed count matrix
- Output: “biological knowledge”

- The are many scRNA-seq analysis methods, each one try to answer to one (or more) biological question

- Examples of scRNA-seq analysis methods
  - Cell clustering
  - Cell labelling
  - Differential expression analysis
  - …

scRNA-seq count data - Summary

- Raw count table
  - Count how many reads/UMIS are associated to each cell and each gene
  - Rough approximation of true (unknown) gene expression level
  - Affected by technical biases rising from the experimental procedure

- Count preprocessing
  - Set of procedures to remove/mitigate technical biases

- Pre-processed count table
  - Output of pre-processing step
  - Better approximation of the true (unknown) gene expression levels

- Analysis
  - Set of procedures to extract biological information from (pre-processed) data

Raw count table

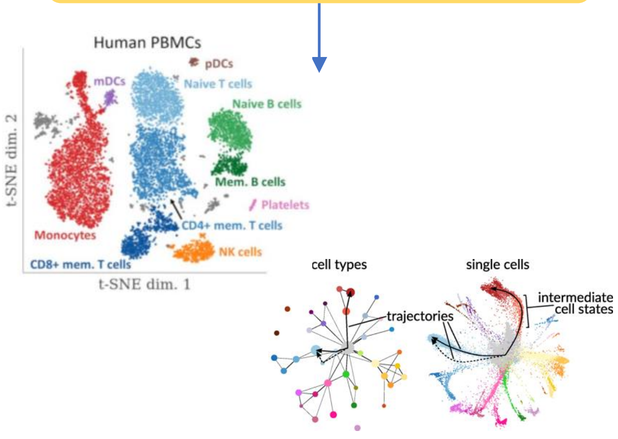
Gene 1	10	9	10		7
Gene 2	3	2	4		0
Gene 3	0	0	0		0
…				…	
Gene N	42	40	37		5

Count pre-processing

Pre-processed count table

Gene 1	10	9	10		7
Gene 2	3	2	4		0
Gene 3	0	0	0		0
…				…	
Gene N	42	40	37		5

Analysis



Why simulating data?

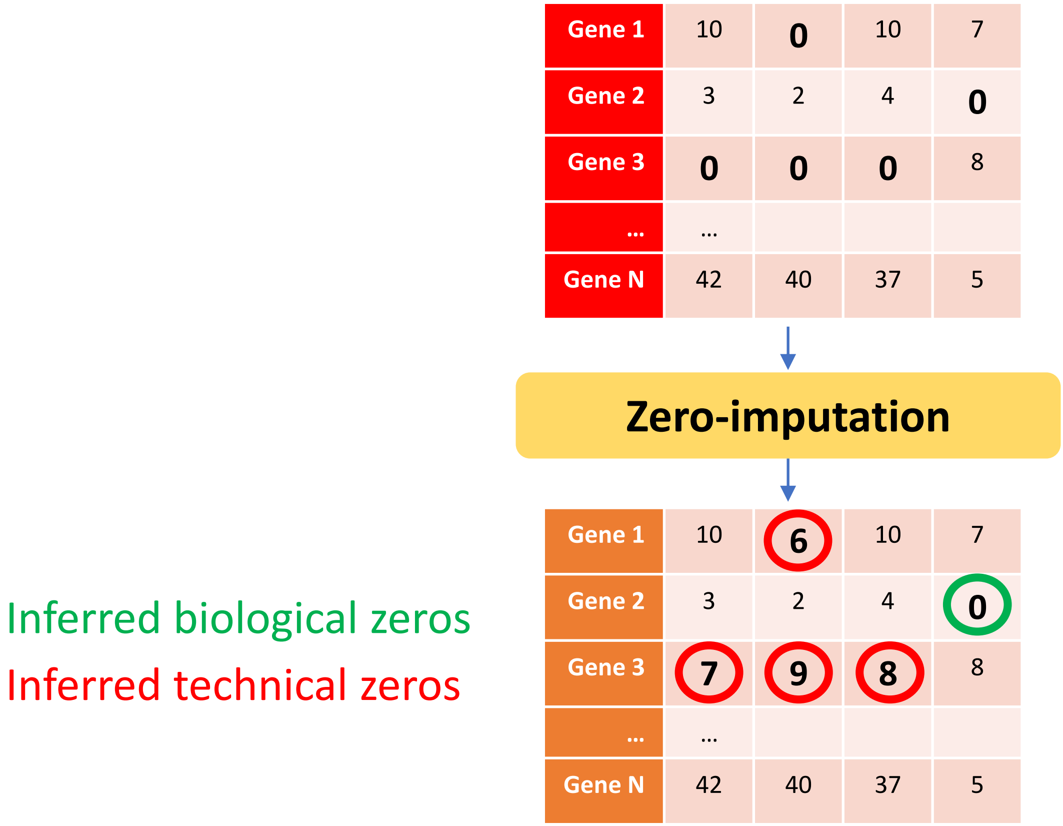
- In future lessons you will see how
  - Pre-process data
  - Analyze data

- In the rest of this lesson, you will see how to simulated scRNA-seq raw count data

- Why simulating raw count data?
- Simulating data allows you to have
  - The true (simulated) gene expression level -> the information that is unknown in real scRNA-seq experiments
  - The corresponding raw count matrix

Zero-imputation

- Goal: fix the technical zeros
- Input: raw count matrix
- Output: imputed count matrix
- How it works:
  - Identify genes having zero values
  - Identify which of these zero values are “technical zeros”
  - Recover (only) the technical zeros



Gene 1	10	0	10	7
Gene 2	3	2	4	0
Gene 3	0	0	0	8
…	…			
Gene N	42	40	37	5

Zero-imputation

Gene 1	10	6	10	7
Gene 2	3	2	4	0
Gene 3	7	9	8	8
…	…			
Gene N	42	40	37	5

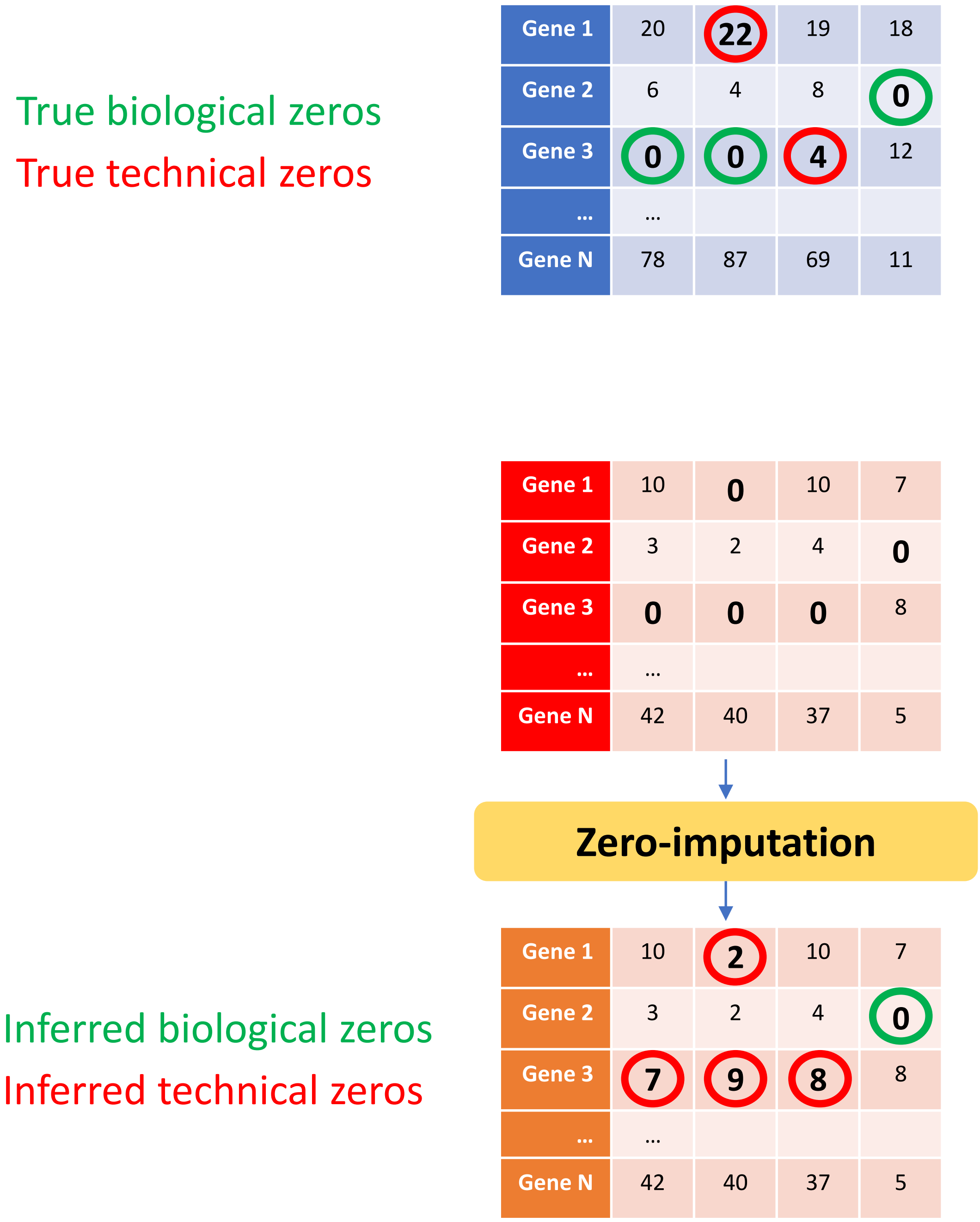
Inferred biological zeros

Inferred technical zeros

Zero-imputation

- Goal: fix the technical zeros
- Input: raw count matrix
- Output: imputed count matrix
- How it works:
    - Identify genes having zero values
    - Identify which of these zero values are “technical zeros”
    - Recover (only) the technical zeros

- Having a ground truth is now possible:
    - Know which zeros are “biological” and which ones are “technical”
    - Know the original gene expression value



True biological zeros

True technical zeros

Gene 1	20	22	19	18
Gene 2	6	4	8	0
Gene 3	0	0	4	12
…	…			
Gene N	78	87	69	11

Gene 1	10	0	10	7
Gene 2	3	2	4	0
Gene 3	0	0	0	8
…	…			
Gene N	42	40	37	5

Zero-imputation

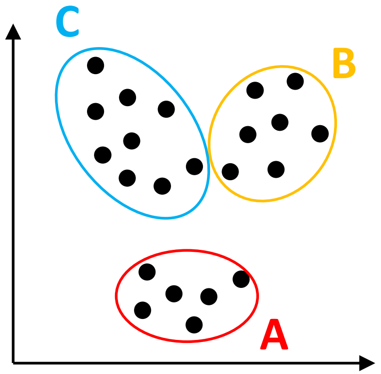
Gene 1	10	2	10	7
Gene 2	3	2	4	0
Gene 3	7	9	8	8
…	…			
Gene N	42	40	37	5

Inferred biological zeros

Inferred technical zeros

Cell-clustering

- Goal: identify group of cells belonging to a common condition (e.g. cell type)
- Input: pre-processed count matrix
- Output: list of cell-group associations
- How it works:
    - Select “important genes”
    - Apply dimensional reduction
    - Cluster cells (in the dimensionally reduced space)

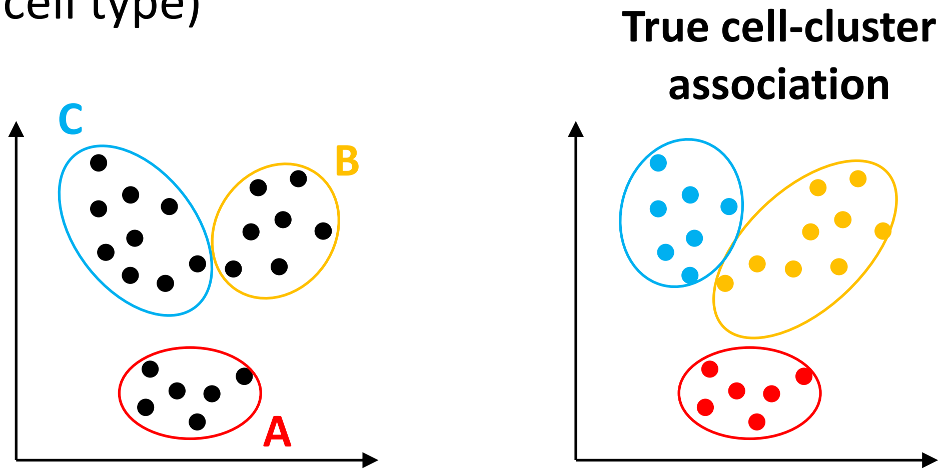


Cell	Inferred cluster
Cell 1	Cluster A
Cell 2	Cluster C
Cell 3	Cluster C
…	…
Cell M	Cluster B

Cell-clustering

- Goal: identify group of cells belonging to a common condition (e.g. cell type)
- Input: pre-processed count matrix
- Output: list of cell-group associations
- How it works:
    - Select "important genes"
    - Apply dimensional reduction
    - Cluster cells (in the dimensionally reduced space)

- Having a ground truth is now possible:
    - Know the true cell-cluster association



Cell	Inferred cluster	True cluster	
Cell 1	Cluster A	Cluster A	✓
Cell 2	Cluster C	Cluster B	✗
Cell 3	Cluster C	Cluster C	✓
…	…		
Cell M	Cluster B	Cluster B	✓

Differential abundance analysis

- Goal: identify genes that are differentially expressed across different conditions
- Input:
  - Pre-processed count matrix
  - Groups of cells
- Output:
  - List of genes detected as DE
  - Optional: “level of confidence”
- How it works:
  - Compare the expression value of each gene across the two conditions
  - Define if there is any difference, the “magnitude” and the “significance”

Gene	Inferred DE
Gene 1	DE
Gene 2	NOT DE
Gene 3	NOT DE
…	…
Gene M	DE

Differential abundance analysis

- Goal: identify genes that are differentially expressed across different conditions
- Input:
  - Pre-processed count matrix
  - Groups of cells
- Output:
  - List of genes detected as DE
  - Optional: “level of confidence”
- How it works:
  - Compare the expression value of each gene across the two conditions
  - Define if there is any difference, the “magnitude” and the “significance”
- Having a ground truth is now possible:
  - Know the true list of DE genes

Gene	Inferred DE	True DE	
Gene 1	DE	DE	✓
Gene 2	NOT DE	DE	✗
Gene 3	NOT DE	NOT DE	✓
…	…		
Gene M	DE	NOT DE	✗

Why simulating - Summary

- Bioinformatics pre-processing and analysis methods/software implements reasonable solutions to complex problems…
- … but scRNA-seq data are hard to handle
    - Large biological variability
    - Many technical biases

- Simulated data provides a way to test/assess the performance, robustness and reliability of bioinformatics methods/software
    - During the development of new methods/software
    - To chose the best ones among the already available methods/software

Single cell RNA-seq count data simulator

- The goal of using simulated data is get access to a ground truth (i.e. simulated gene expression level)

- scRNA-seq count data simulator provides as output
    - Simulated gene expression levels (ground truth)
    - Corresponding raw count matrix
    - Additional information (e.g. cell groups)

- A scRNA-seq count data simulator works:
    - Simulating gene expression levels of multiple genes/cells (biological variability)
    - Simulating the experimental procedure to obtain raw counts (technical variability)

Single cell RNA-seq count data simulator

- Available scRNA-seq count data simulators:
  - ***Splatter*** [Zappia et al., Genome Biology, 2017] Article Software
  - ***SymSim*** [Zhang et al., Nature Communications, 2019] Article Software
  - ***SPARSim*** [Baruzzo et al., Bioinformatics, 2019] Article Software
  - …

- We will use ***SPARSim*** to "play" with simulated data:
  - Understand the characteristic of single cell count data
  - Understand how to simulate scRNA-seq count data

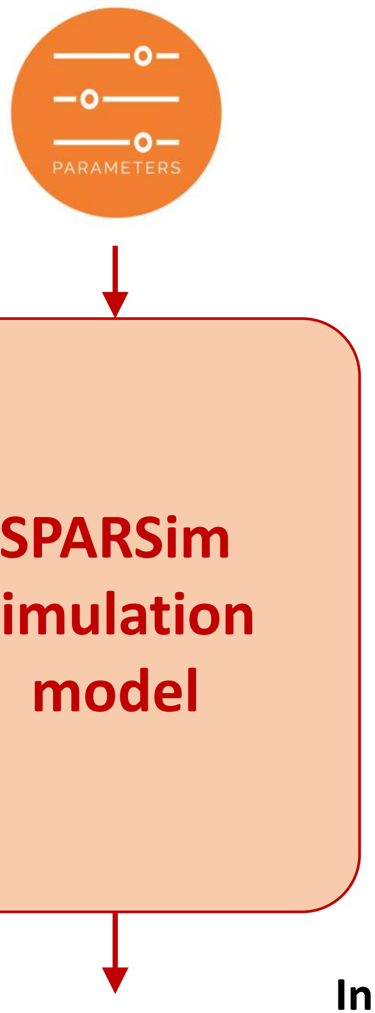
SPARSim

- Article link: https://academic.oup.com/bioinformatics/article
- Software link: https://gitlab.com/sysbiobig/sparsim

- Language: R/C++

- SPARSim can simulate
  - biological and technical variability
  - spike-ins
  - batch effects
  - bimodal gene expression
  - differentially expressed genes
  - multiple cell groups/types

How SPARSim works



SPARSim simulation model

In-silico count table

	Cell_ 1	Cell_ 2	…	Cell_ M
Gene_1	0	12		0
Gene_2	35	0		0
Gene_3	75	128		77
…			…	
Gene_N	0	3		0

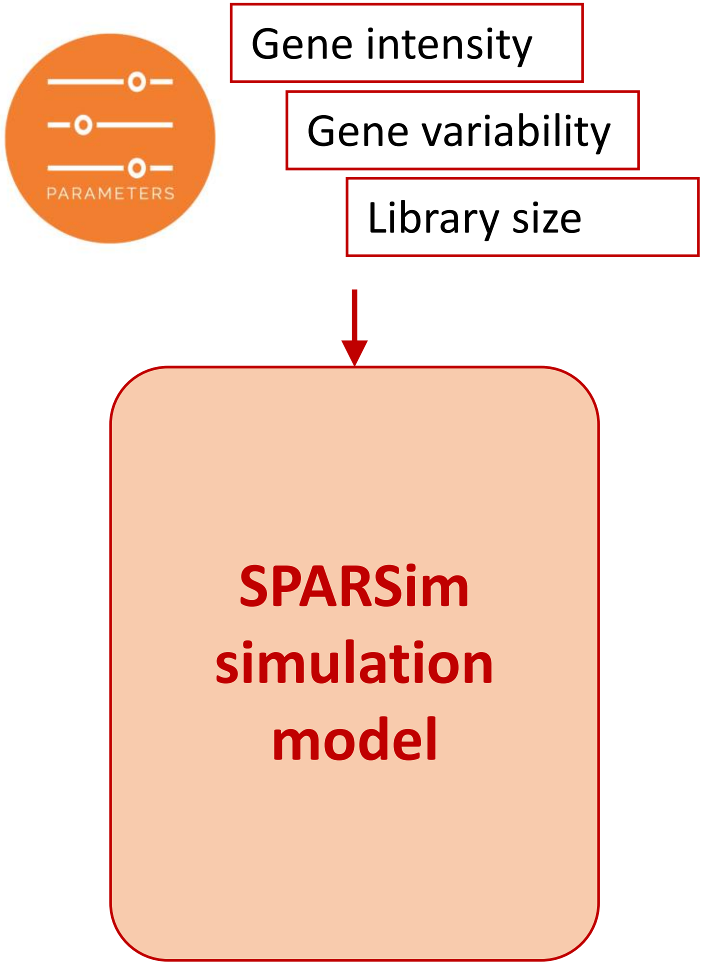
In-silico gene expression level

	Cell_ 1	Cell_ 2	…	Cell_ M
Gene_1	0	2		0
Gene_2	3	2		0
Gene_3	7	13		8
…			…	
Gene_N	1	4		0

How SPARSim works

For each condition to simulate, the user must specify:

- Array of N gene intensity values
- Array of N gene variability values
- Array of M library size values



Gene intensity

Gene variability

Library size

SPARSim simulation model

How SPARSim works

Gamma

$$p(x;k,\theta) = \frac{x^{k-1}e^{-\frac{x}{\theta}}}{\theta^{k}\Gamma(k)}$$

$$x, k, \theta > 0$$

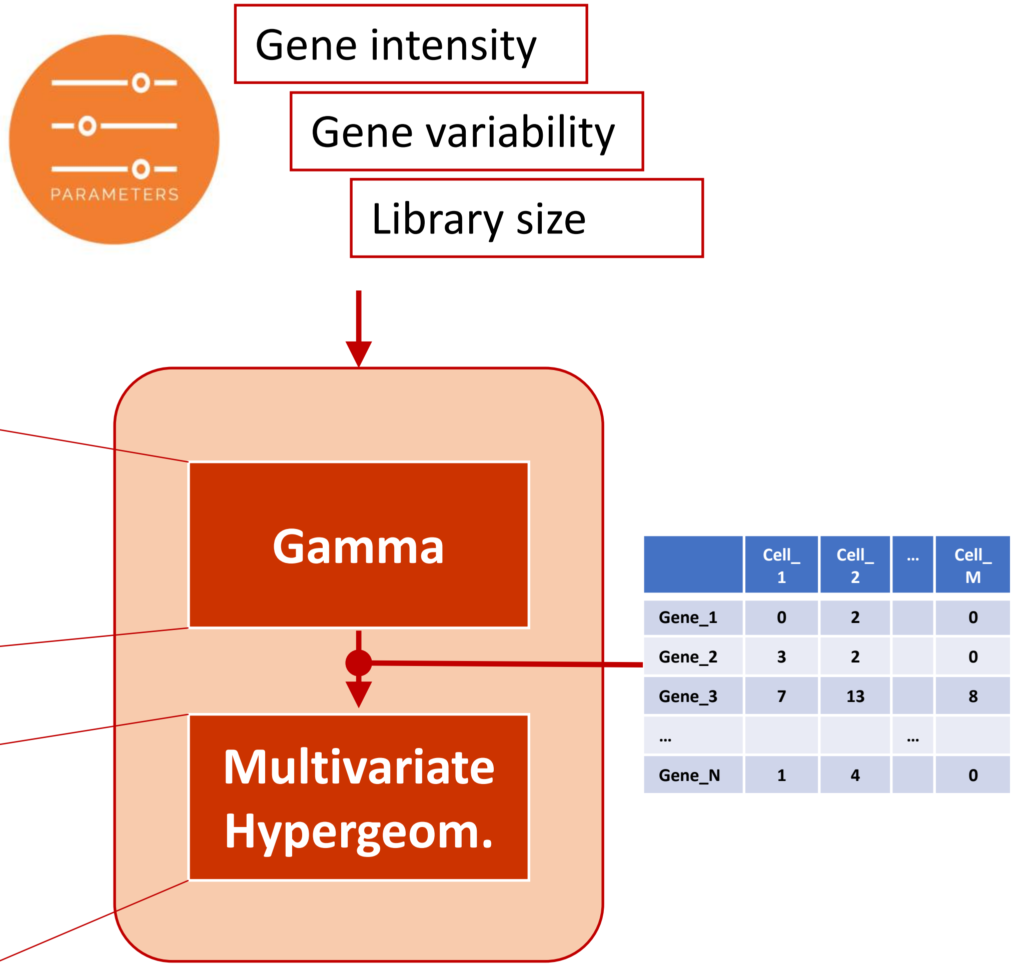
$$\Gamma(k) = \int_{0}^{\infty} y^{k-1}e^{-y}dy$$

Multivariate Hypergeometric

$$p(k_1, k_2, \dots k_c) = \frac{\prod_{i=1}^{c}\binom{K_i}{k_i}}{\binom{N}{n}}$$

$$K_i \in \mathbb{N},\ \sum_i^c K_i = N$$

$$k_i \in \mathbb{N}, \forall i\ k_i < K_i, \sum_i^c k_i = n$$



Gene intensity

Gene variability

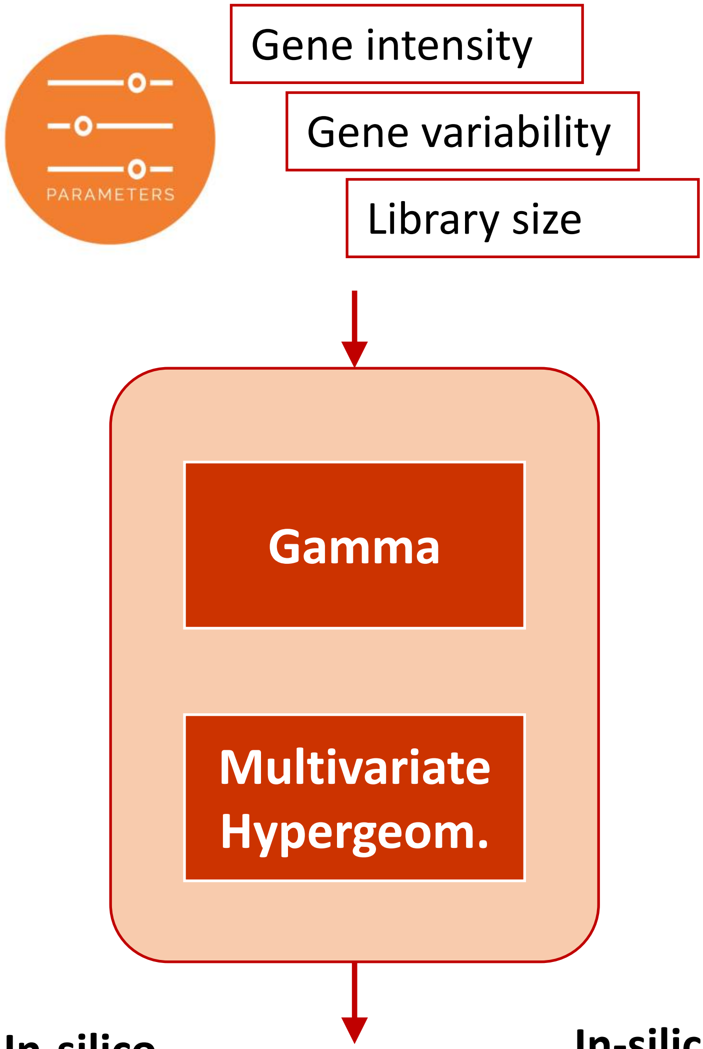
Library size

Gamma

Multivariate Hypergeom.

	Cell_ 1	Cell_ 2	...	Cell_ M
Gene_1	0	2		0
Gene_2	3	2		0
Gene_3	7	13		8
...			...	
Gene_N	1	4		0

How SPARSim works



In-silico count table

	Cell_ 1	Cell_ 2	…	Cell_ M
Gene_1	0	12		0
Gene_2	35	0		0
Gene_3	75	128		77
…			…	
Gene_N	0	3		0

In-silico gene expression level

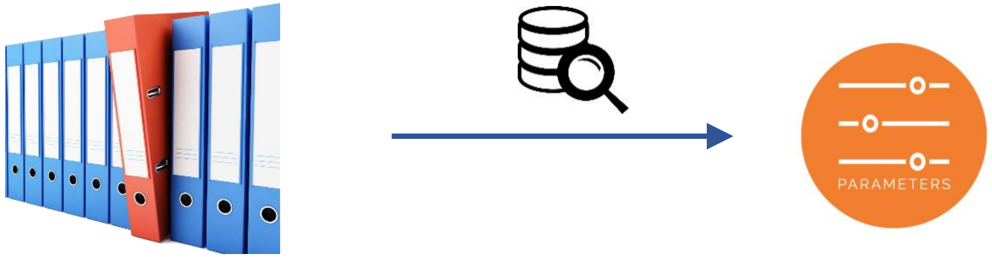
	Cell_ 1	Cell_ 2	…	Cell_ M
Gene_1	0	2		0
Gene_2	3	2		0
Gene_3	7	13		8
…			…	
Gene_N	1	4		0

How to run SPARSim

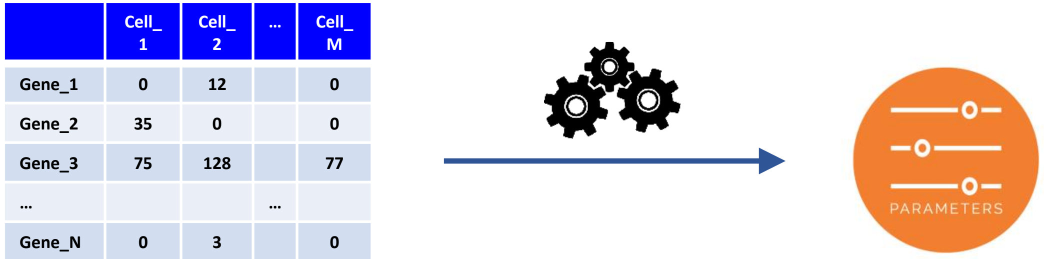
Simulation parameters can be provided in 3 ways:

- Direct input by user

- Using a parameter presets



- Estimation from existing count table
    - SPARSim built-in function



	Cell_ 1	Cell_ 2	…	Cell_ M
Gene_1	0	12		0
Gene_2	35	0		0
Gene_3	75	128		77
…			…	
Gene_N	0	3		0

SPARSim parameters presets

Dataset	Cells	# cells	Sparsity	Platform and protocol	UMIs
Horning et al.	Human LNCap	144	~39%	Smart-seq2	No
Engel et al.	Mouse NK T cells	203	~69%	Flow cytometry, Smart-seq2*	No
Tung et al.	Human IPSCs	564	~53%	Fluidigm C1, SMARTer*	Yes
Camp et al.	Human Brain/cerebral organoids	434	~84%	Fluidigm C1, SMARTer	No
Chu et al.	Human iPSC to Endoderm	758	~51%	Fluidigm C1, SMARTer	No
Bacher et al.	Human ESCs	366	~39%	Fluidigm C1, SMARTer	No
10x PBMC	Human Peripheral blood mononuclear cells	5419	~96%	10x Genomics	Yes
10x T	Human Pan T cells	8093	~95%	10x Genomics	Yes
10x Brain	Mouse Brain cells	11843	~87%	10x Genomics	Yes
Zheng et al.	Human Jurkat and 293T cells	3388	~83%	10x Genomics	Yes
Macosko et al.	Mouse Retinal cells	9000	~97%	Drop-Seq	Yes
Saunders et al.	Mouse Polydendrocytes cells	5688	~94%	Drop-Seq	Yes

SPARSim – Create simulation parameter

Function to create the simulation parameter describing one condition

Documentation: **?SPARSim_create_simulation_parameter**

```
SPARSim_create_simulation_parameter(
      intensity,
      variability,
      library_size,
      feature_names = NA,
      sample_names = NA,
      condition_name = NA
)
```

SPARSim – Create simulation parameter

The return value of the function is a list of 4 elements:

- **`..$intensity`**
- **`..$variability`**
- **`..$lib_size`**
- **`..$name`**

intensity	variability	library_ size	name
…	…	…	…
…	…	…	
…	…	…	
…	…	…	
…	…	…	
…	…	…	
…	…		
…	…		

The function packs all the information about the genes/cells to simulate in a list of 4 elements

Example

```
# Simulate
# - 1 condition
# - 8 genes
# - 300 cells

gene_int <- c(  80,  30,   15,  10,  10,   5,   3,   0)
gene_var <- c(0.05, 0.1,  0.1, 0.2, 0.5, 1.0, 5.0, 1.0)
gene_IDs <- c("Gene_1", "Gene_2" ,"Gene_3", "Gene_4", "Gene_5", "Gene_6", "Gene_7", "Gene_8")


# Number of cells to simulated
N_cell <- 300

# Simulate 300 cells with a constant library size of 50
lib_size <- rep(50, times = N_cell)

# create simulation parameter
cond_A_param <- SPARSim_create_simulation_parameter(
                            intensity =  gene_int ,
                            variability = gene_var,
                            library_size = lib_size,
                            feature_names = gene_IDs,
                            sample_names = paste0("cond_A_cell_",c(1:N_cell)),
                            condition_name = "condition_A")
```

SPARSim – Create DE genes

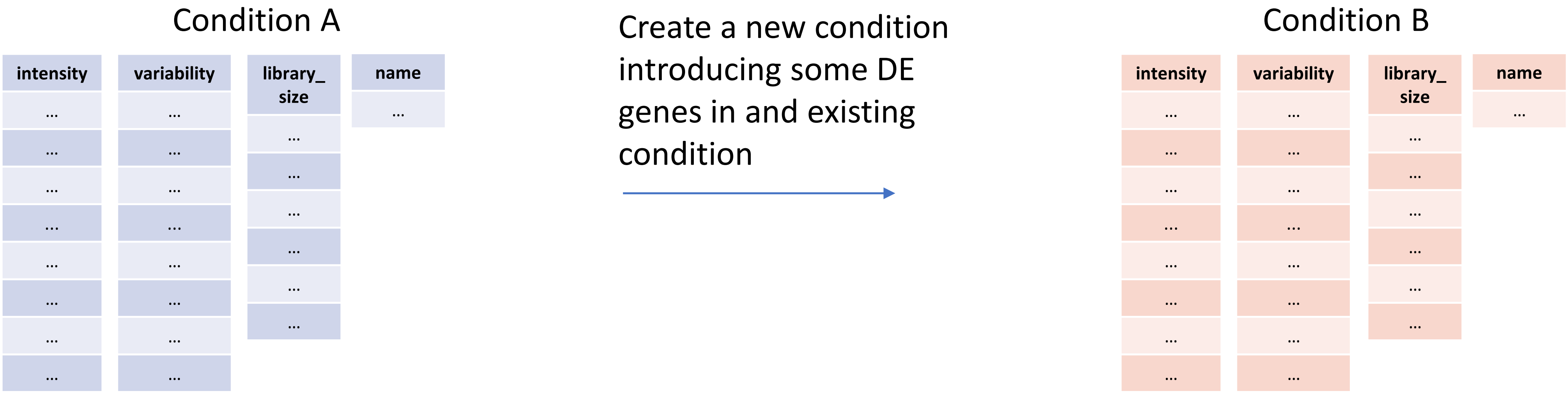
Function to create a new condition starting from an existing one, just introducing DE genes

Documentation: **`?SPARSim_create_DE_genes_parameter`**

```
SPARSim_create_DE_genes_parameter(
      sim_param,
      fc_multiplier,
      N_cells = NULL,
      lib_size_DE = NULL,
      sample_names = NULL,
      condition_name = NULL
)
```

SPARSim – Create DE genes

Idea:



Condition A

intensity	variability	library_ size	name
…	…	…	…
…	…	…	
…	…	…	
…	…	…	
…	…	…	
…	…	…	
…	…		
…	…		

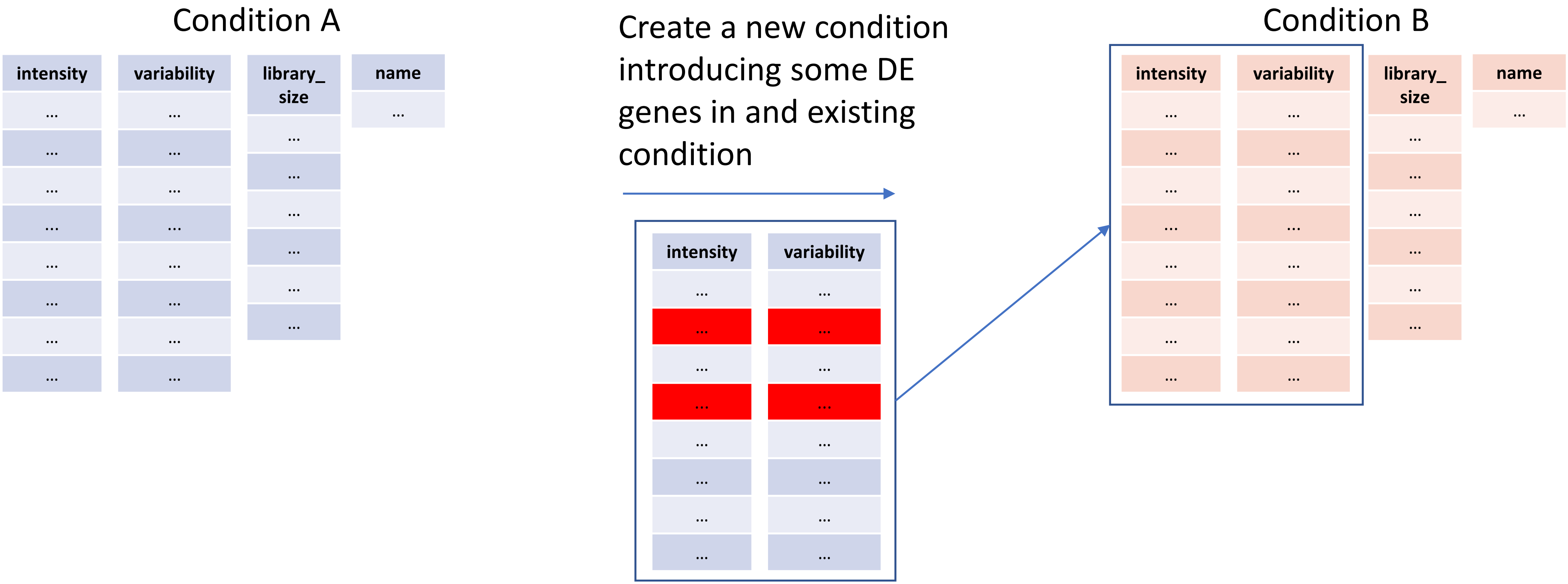
Create a new condition introducing some DE genes in and existing condition

Condition B

intensity	variability	library_ size	name
…	…	…	…
…	…	…	
…	…	…	
…	…	…	
…	…	…	
…	…	…	
…	…		
…	…		

SPARSim – Create DE genes

Idea:



Condition A

intensity	variability	library_ size	name
...	...	...	...
...	...	...	
...	...	...	
...	...	...	
...	...	...	
...	...	...	
...	...		
...	...		

Create a new condition introducing some DE genes in and existing condition

intensity	variability
...	...
...	...
...	...
...	...
...	...
...	...
...	...
...	...

Condition B

intensity	variability	library_ size	name
...	...	...	...
...	...	...	
...	...	...	
...	...	...	
...	...	...	
...	...	...	
...	...		
...	...		

Example

```
# Create “condition B” starting from “condition A”
# Compared with “condition A”, the new “condition B” will have
#  - 3 DE genes
#  - 500 cells

cond_A_param # simulation parameter of “condition A”

# Gene 1: increase 4 times its expression level
# Gene 2: decrease half its expression level
# Gene 3: increase 2 times its expression level
# Other genes: keep their original expression levels
DE_multiplier <- c(4, 0.5, 2, 1, 1, 1, 1, 1, 1)

# create simulation parameter
cond_B_param <- SPARSim_create_DE_genes_parameter(
                                 sim_param = cond_A_param,
                                 fc_multiplier = DE_multiplier,
                                 N_cells = 500,
                                 condition_name = "condition_B")
```

SPARSim - Simulate

Function to simulate gene expression level and the corresponding count matrix

Documentation: **`?SPARSim_simulation`**

```
SPARSim_simulation(
                   dataset_parameter
                   )
```

Example

```
# Collect the previously created simulation parameters
sim_param <- list(cond_A_param, cond_B_param) # 2 conditions

# Simulate the two conditions
sim_results <- SPARSim_simulation(dataset_parameter = sim_param)
```

sim_results is a list of several elements, the most important are

- **`..$gene_matrix`**: simulated gene expression level
- **`..$count_matrix`**: simulated raw count matrix

Let’s simulate scRNA-seq data!

Do the simulation in R
Inspect the simulated data with R and/or Python

Simulation examples:

- **Example 1**: toy example (1 condition, equal library size)
- **Example 2**: toy example (effect of sequencing depth, technical zeros)
- **Example 3**: toy example (effect of uneven sequencing depth, role of normalization)
- **Example 4**: realistic example (from parameters preset)
- **Example 5**: create multiple conditions and DE genes

Jupyter notebook available on git: https://github.com/baruz89/scRNA-seq_simulation_Webvalley2020

Simulation example 1

Description: first toy example

- 1 condition
- 8 genes (skewed internal distribution)
- 300 cells
- equal library size (ideal)

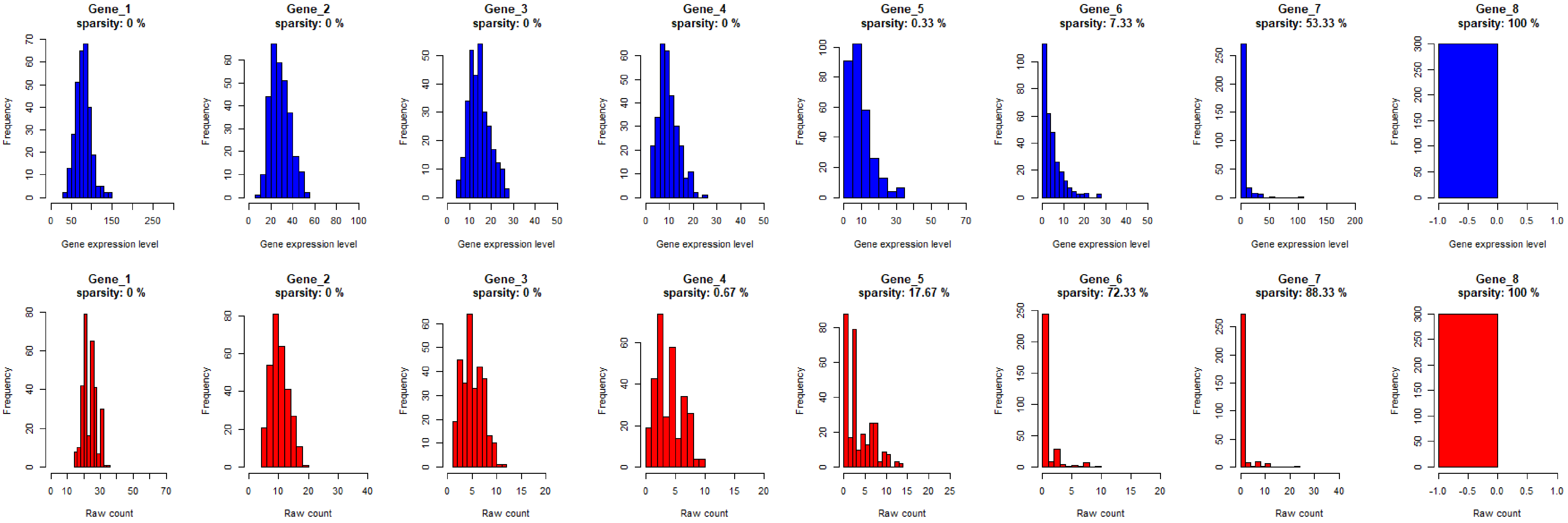
Gene	Intensity
Gene 1	80
Gene 2	30
Gene 3	15
Gene 4	10
Gene 5	10
Gene 6	5
Gene 7	3
Gene 8	0

Gene	Variability
Gene 1	0.05
Gene 2	0.1
Gene 3	0.1
Gene 4	0.2
Gene 5	0.5
Gene 6	1.0
Gene 7	5.0
Gene 8	1.0

Gene	library_size
Cell 1	50
Cell 2	50
Cell 3	50
…	…
Cell 300	50

Code: https://github.com/baruz89/scRNA-seq_simulation_Webvalley2020/tree/master/Example%201

Simulation example 1



Simulation example 2

Description: toy example, effect of sequencing depth, technical zeros

- 1 condition
- 8 genes (skewed internal distribution)
- 300 cells
- 3 levels of sequencing depth (high, medium, low)

Code: https://github.com/baruz89/scRNA-seq_simulation_Webvalley2020/tree/master/Example%202

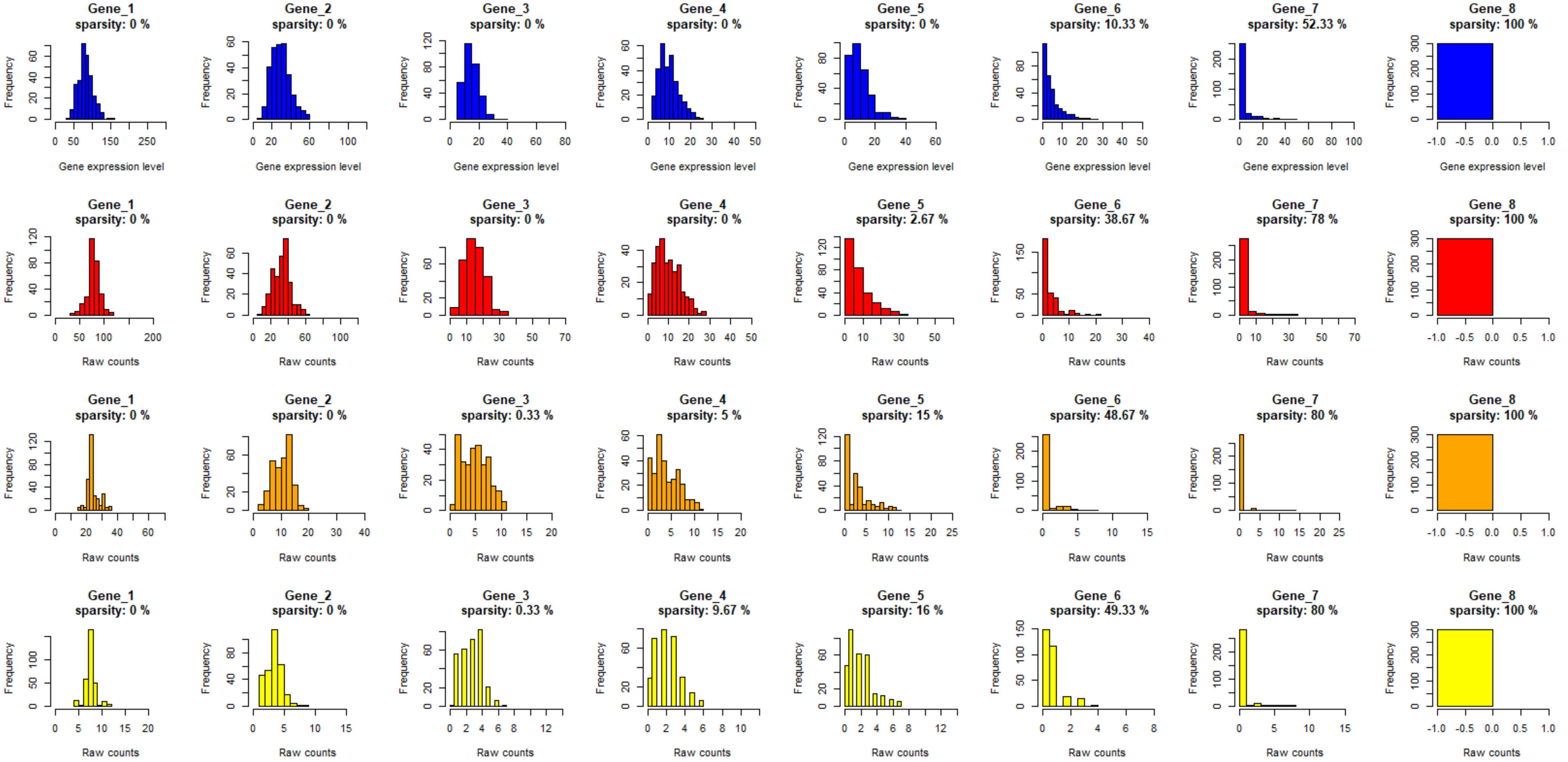
Simulation example 2

Gene expression levels

High sequencing depth

Medium sequencing depth

Low sequencing depth



Simulation example 3

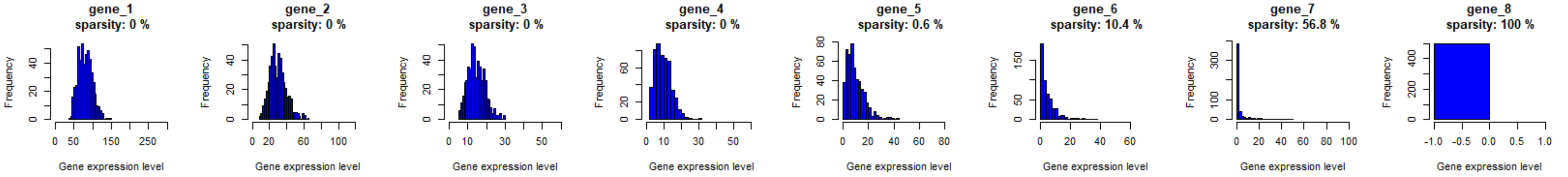
Description: toy example, uneven sequencing depth

- 1 conditions
- 8 genes (skewed internal distribution)
- 500 cells
- uneven library size

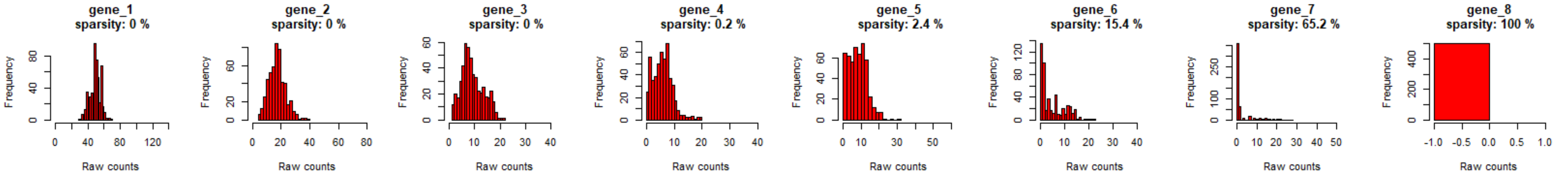
Code: https://github.com/baruz89/scRNA-seq_simulation_Webvalley2020/tree/master/Example%203

Simulation example 3

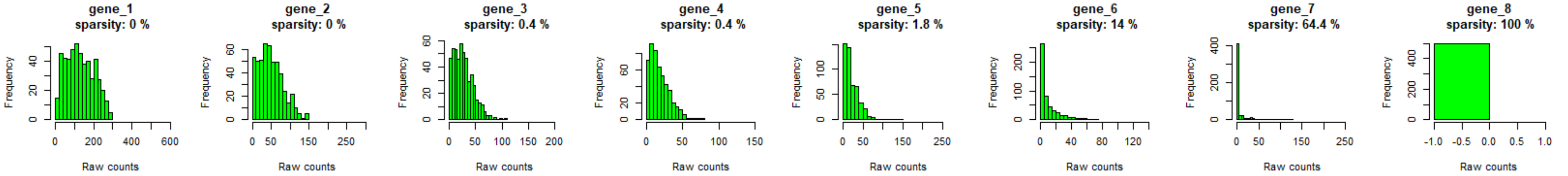
Gene expression levels



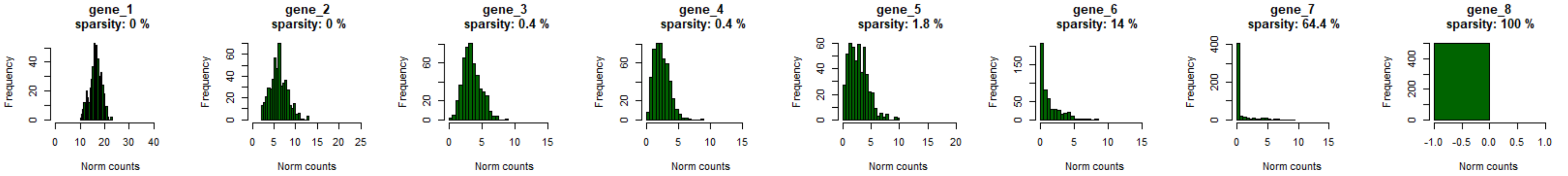
Even sequencing depth



Uneven sequencing depth



Uneven sequencing depth + normalization



Simulation example 4

Description: use parameter preset

- 4 conditions
- ~19K genes
- ~4K cells

Code: https://github.com/baruz89/scRNA-seq_simulation_Webvalley2020/tree/master/Example%204

Simulation example 4

?Zheng_param_preset

..$Zheng_C1

Intensity	Variability	library_ size
…	…	…
…	…	…
…	…	…
…	…	…
…	…	…
…	…	…
…	…	
…	…	

- 19536 genes
- 1440 cells
- Jurkat cells

..$Zheng_C2

Intensity	Variability	library_ size
…	…	…
…	…	…
…	…	…
…	…	…
…	…	…
…	…	…
…	…	
…	…	

- 19536 genes
- 1718 cells
- T cells

..$Zheng_C3

Intensity	Variability	library_ size
…	…	…
…	…	…
…	…	…
…	…	…
…	…	…
…	…	…
…	…	
…	…	

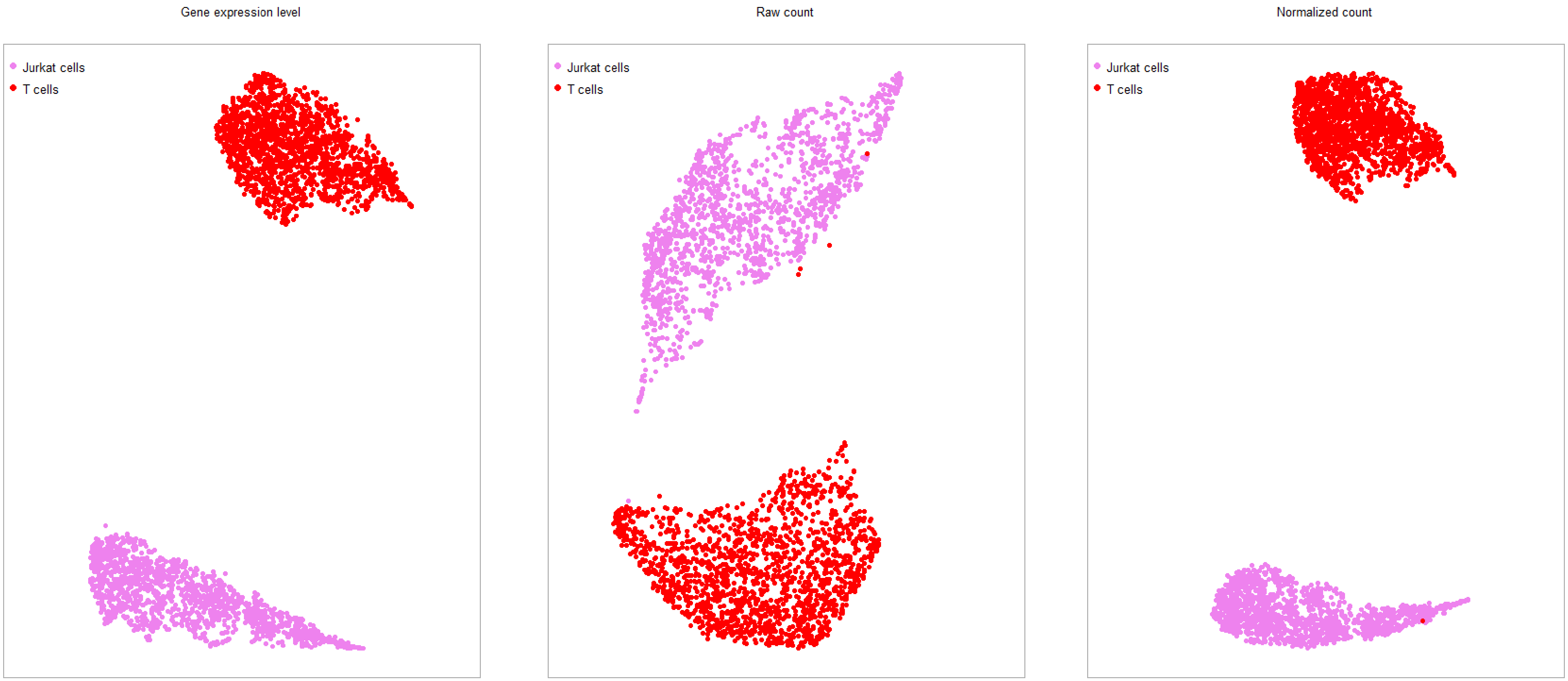
- 19536 genes
- 184 cells
- Mix cells

..$Zheng_C4

Intensity	Variability	library_ size
…	…	…
…	…	…
…	…	…
…	…	…
…	…	…
…	…	…
…	…	
…	…	

- 19536 genes
- 46 cells
- Unknown cells

Simulation example 4



Gene expression level

Raw count

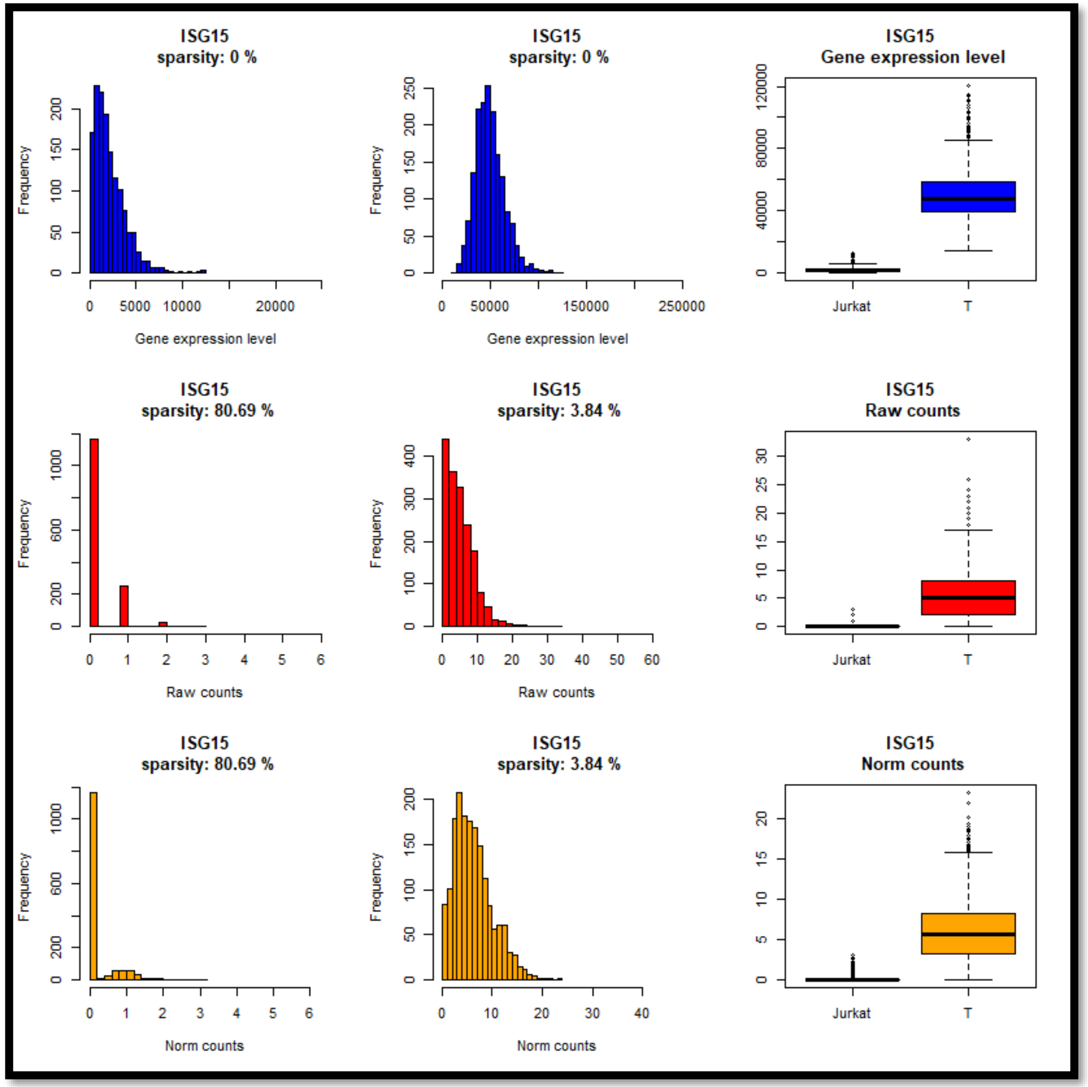
Normalized count

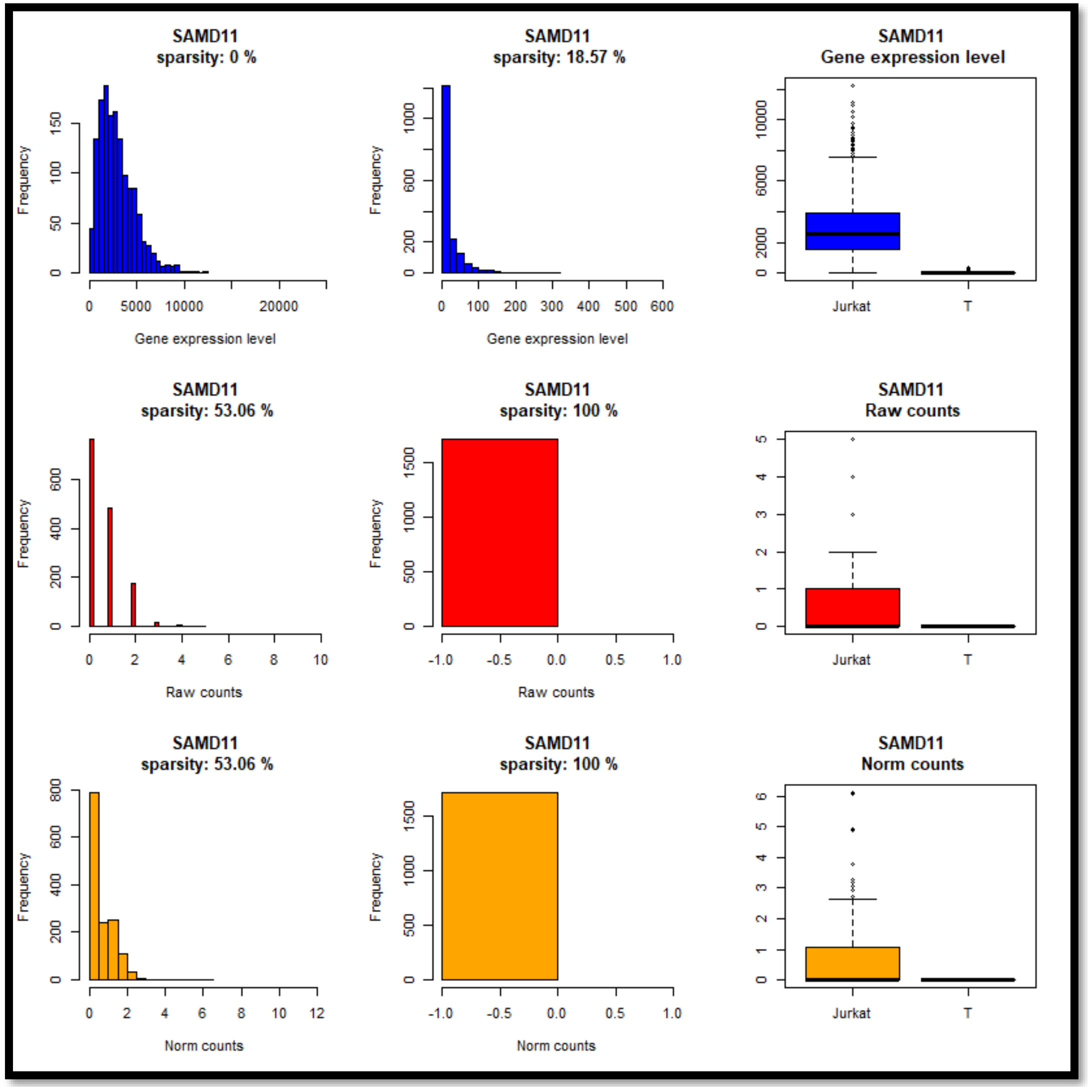
Simulation example 4

Gene expression levels

Raw count

Pre-processed count



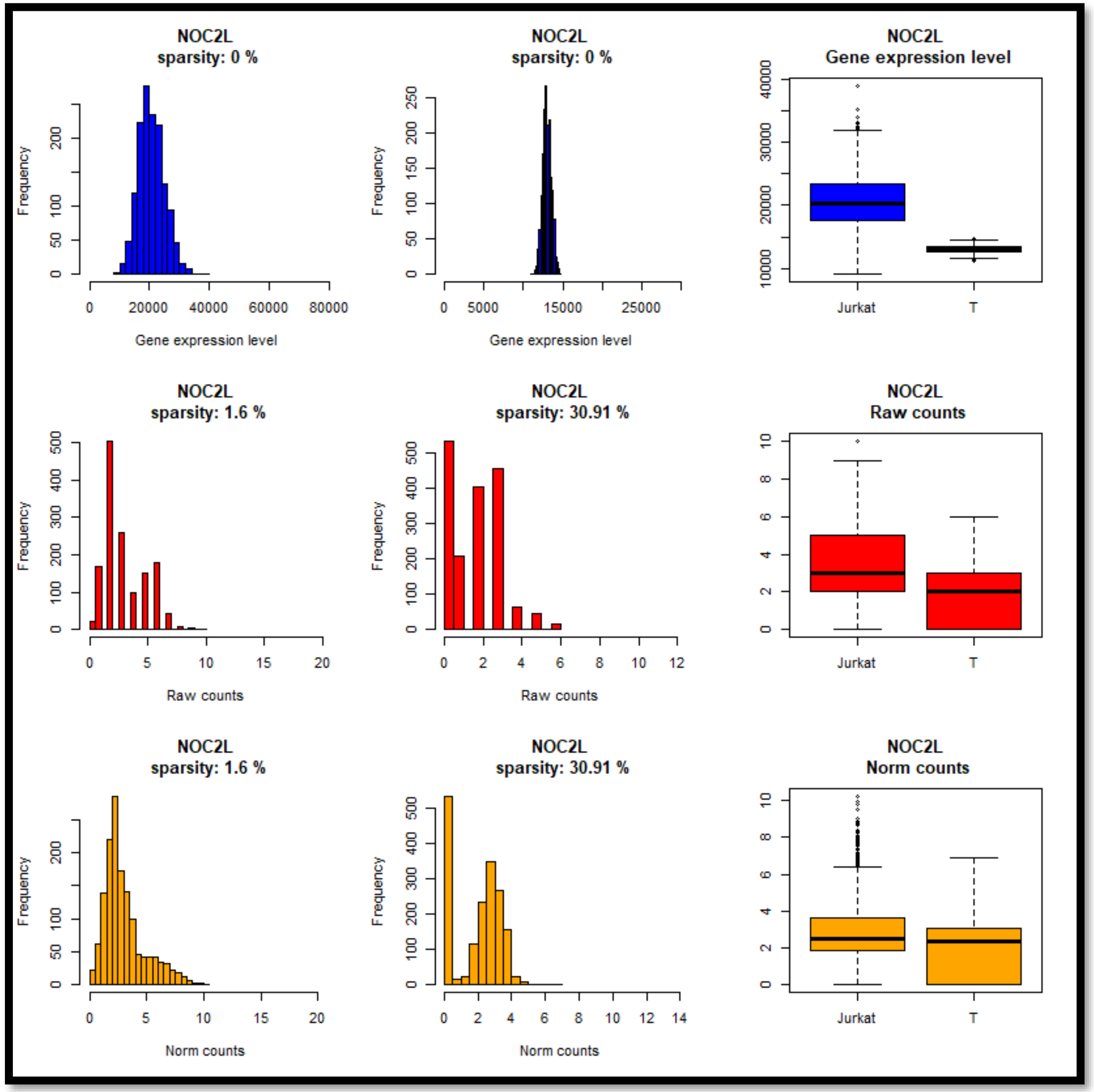


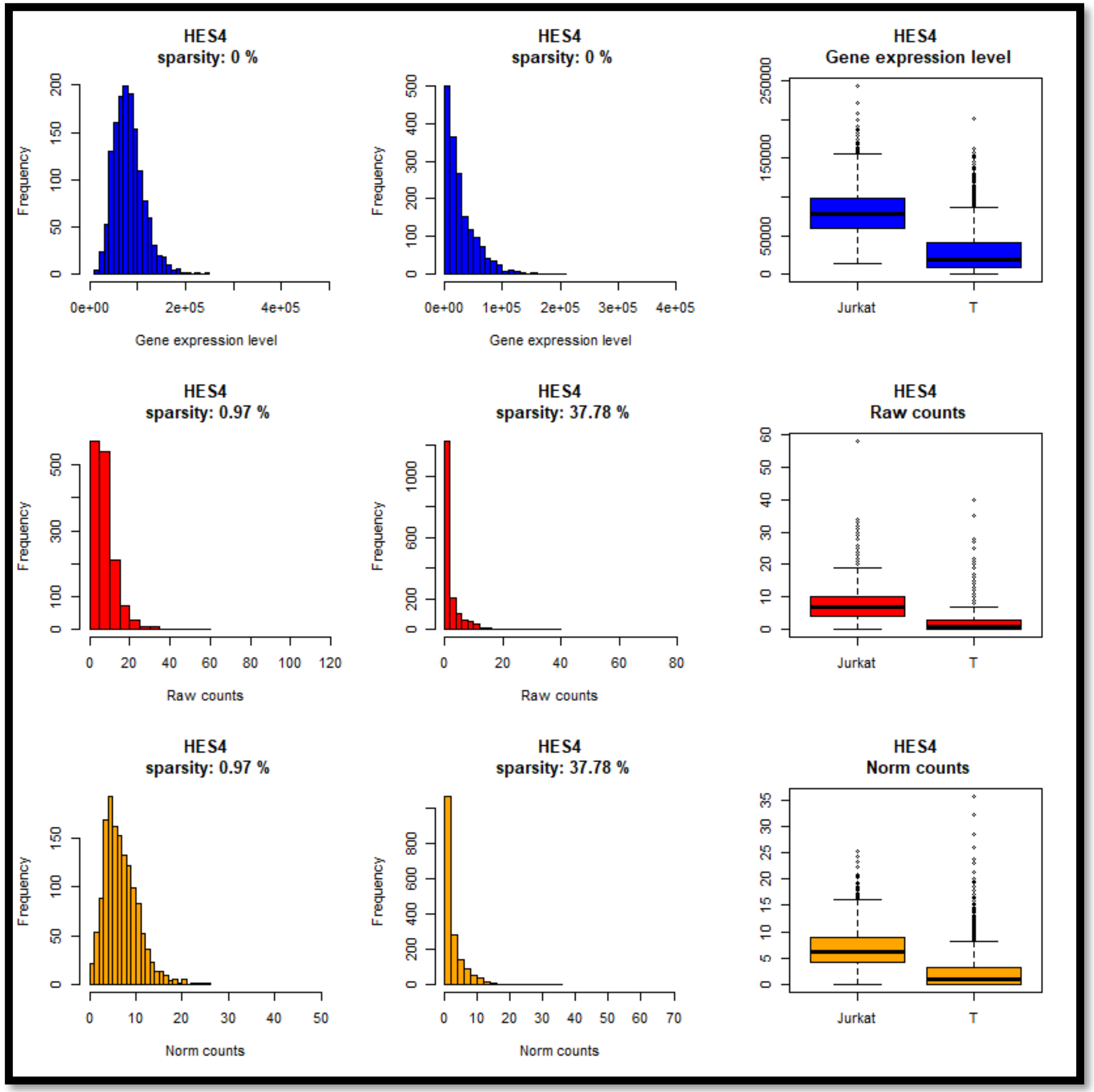
Simulation example 4

Gene expression levels

Raw count

Pre-processed count



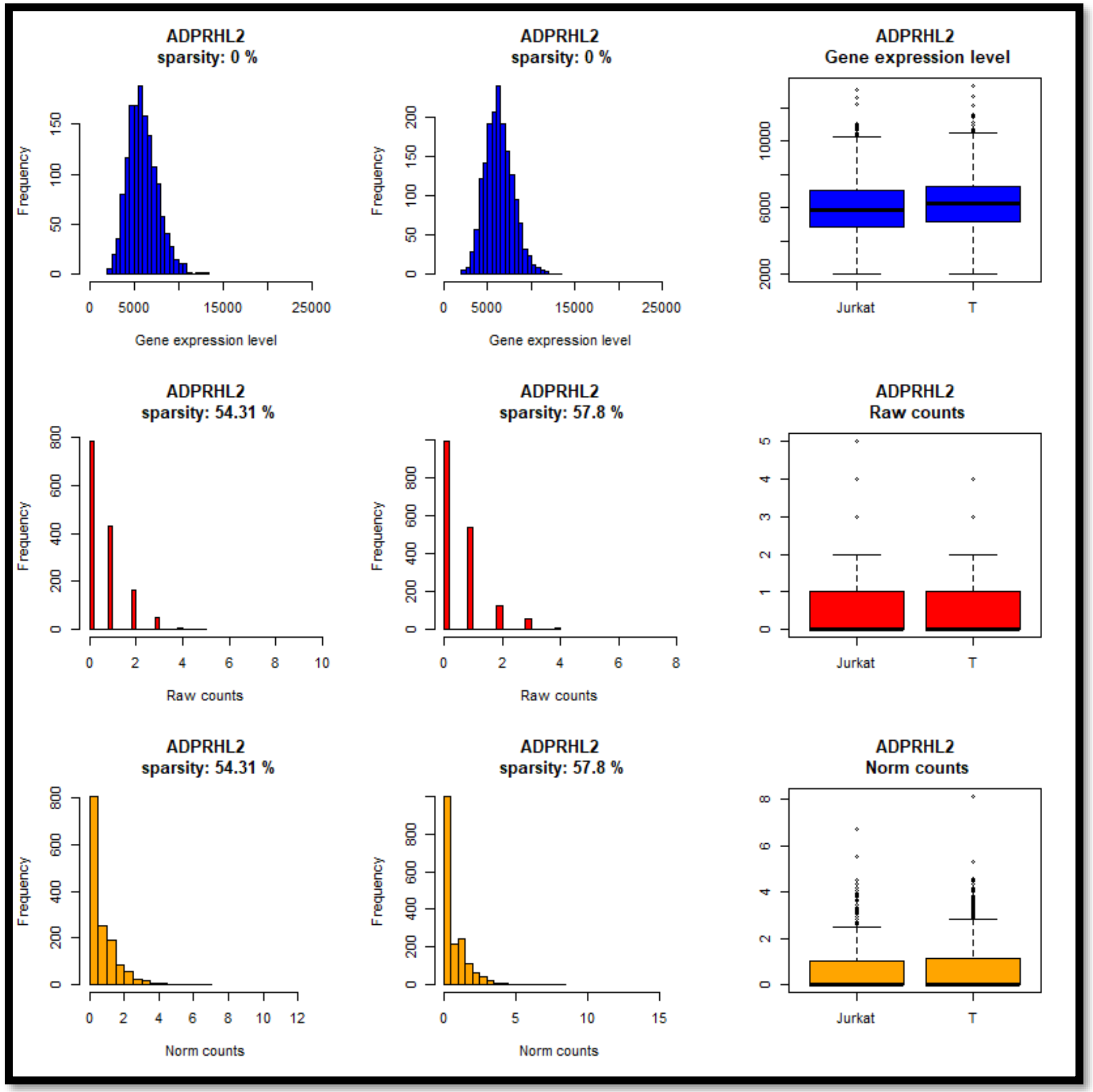


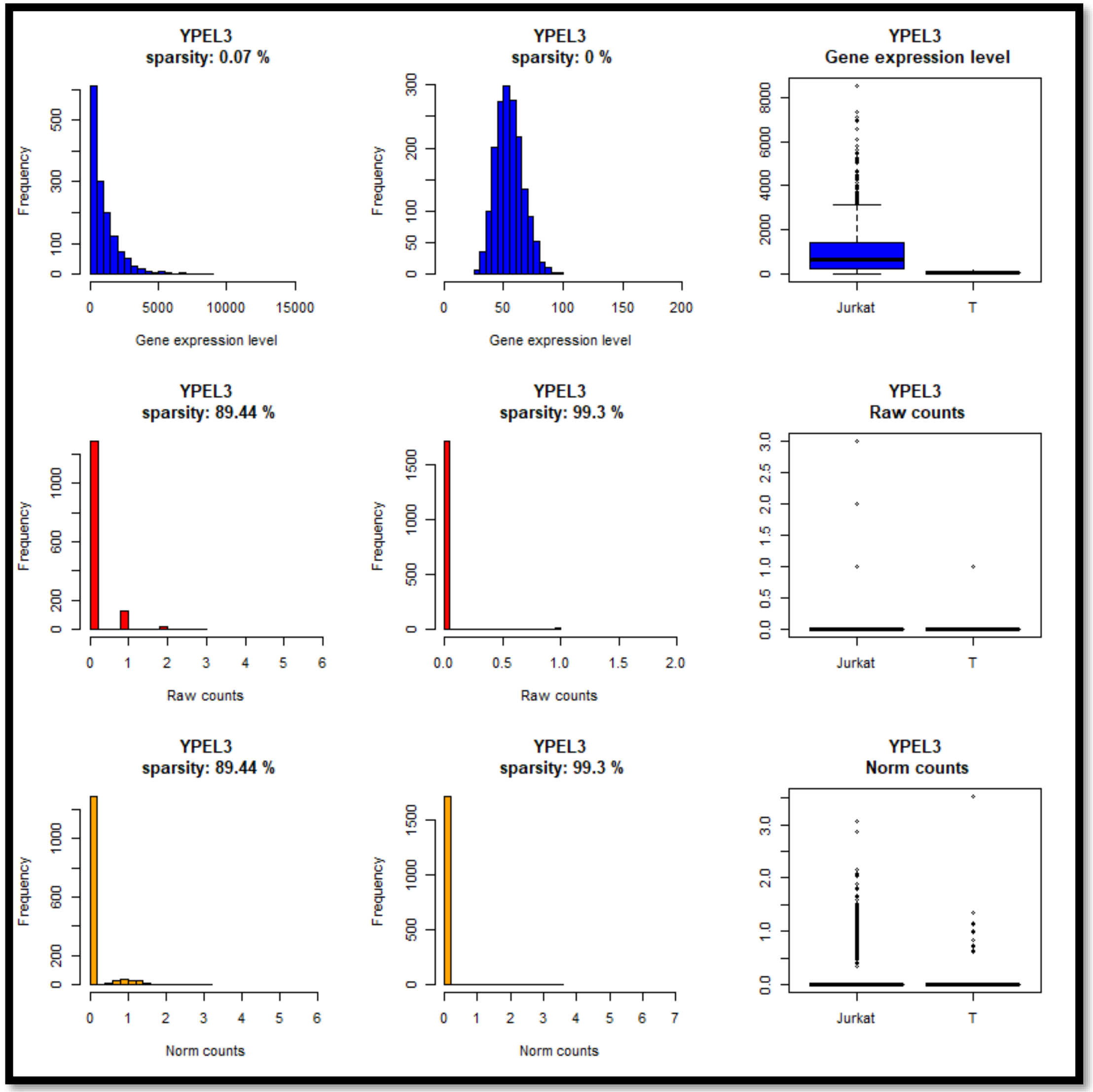
Simulation example 4

Gene expression levels

Raw count

Pre-processed count





Simulation example 5

Description: extend a parameter preset, create new conditions, create DE genes

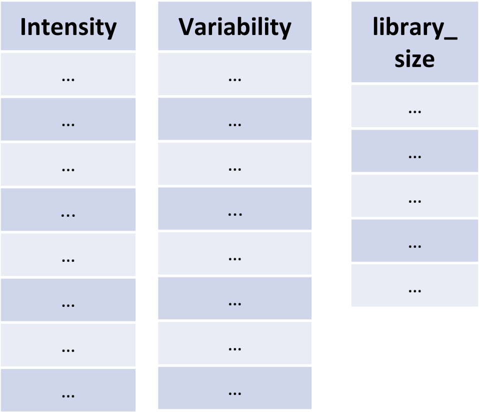
- 2 conditions -> 5 conditions
- ~19K genes
- ~1K cells
- DE genes

Code: https://github.com/baruz89/scRNA-seq_simulation_Webvalley2020/tree/master/Example%205

Simulation example 5

?Zheng_param_preset

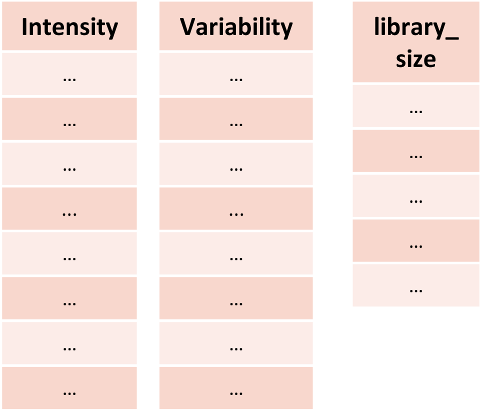
..$Zheng_C1



Intensity	Variability	library_size
…	…	…
…	…	…
…	…	…
…	…	…
…	…	…
…	…	
…	…	
…	…	

- 19536 genes
- ~~1440~~ 200 cells
- Jurkat cells

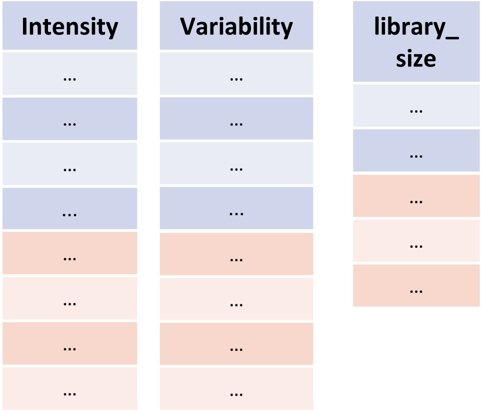
..$Zheng_C2



Intensity	Variability	library_size
…	…	…
…	…	…
…	…	…
…	…	…
…	…	…
…	…	
…	…	
…	…	

- 19536 genes
- ~~1718~~ 200 cells
- T cells

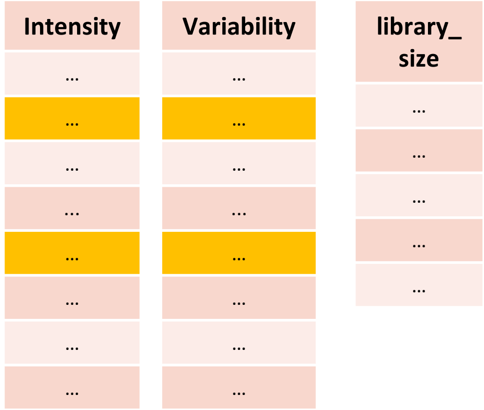
..$Zheng_C1C2_Mix



Intensity	Variability	library_size
…	…	…
…	…	…
…	…	…
…	…	…
…	…	…
…	…	
…	…	
…	…	

- 19536 genes
- 200 cells
- “new” Jurkat&T cells

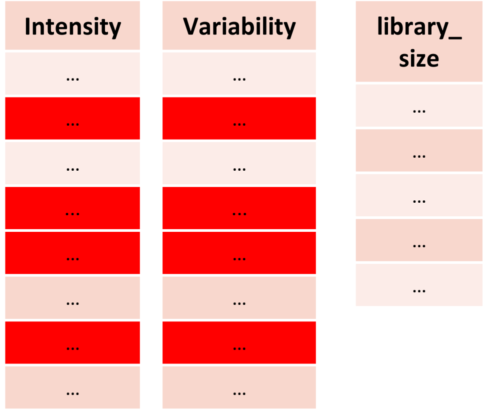
..$Zheng_C2_vA



Intensity	Variability	library_size
…	…	…
…	…	…
…	…	…
…	…	…
…	…	…
…	…	
…	…	
…	…	

- 19536 genes
- 200 cells
- “new” T cells ver. A (few DE genes, small differences)

..$Zheng_C2_vB

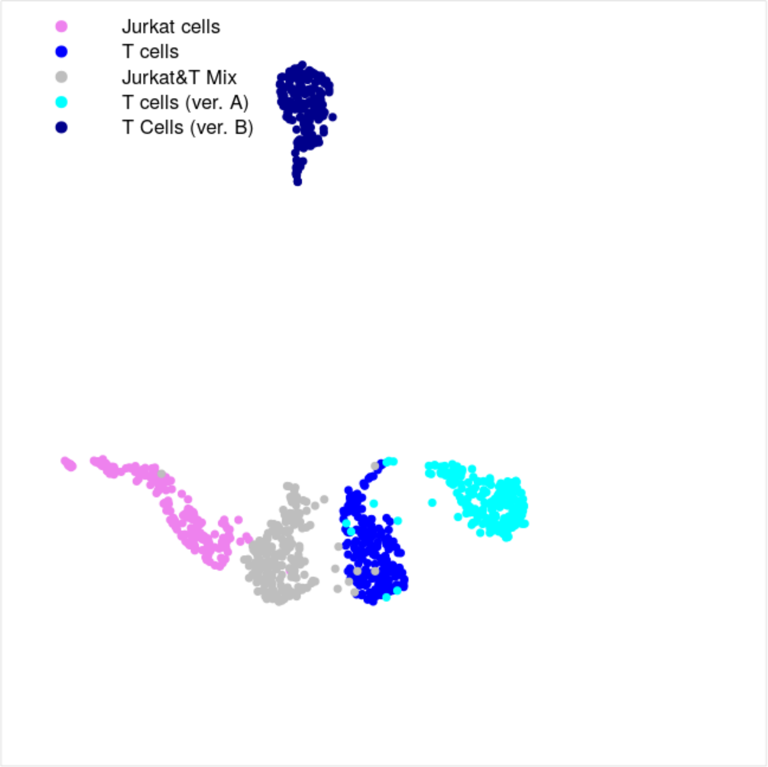


Intensity	Variability	library_size
…	…	…
…	…	…
…	…	…
…	…	…
…	…	…
…	…	
…	…	
…	…	

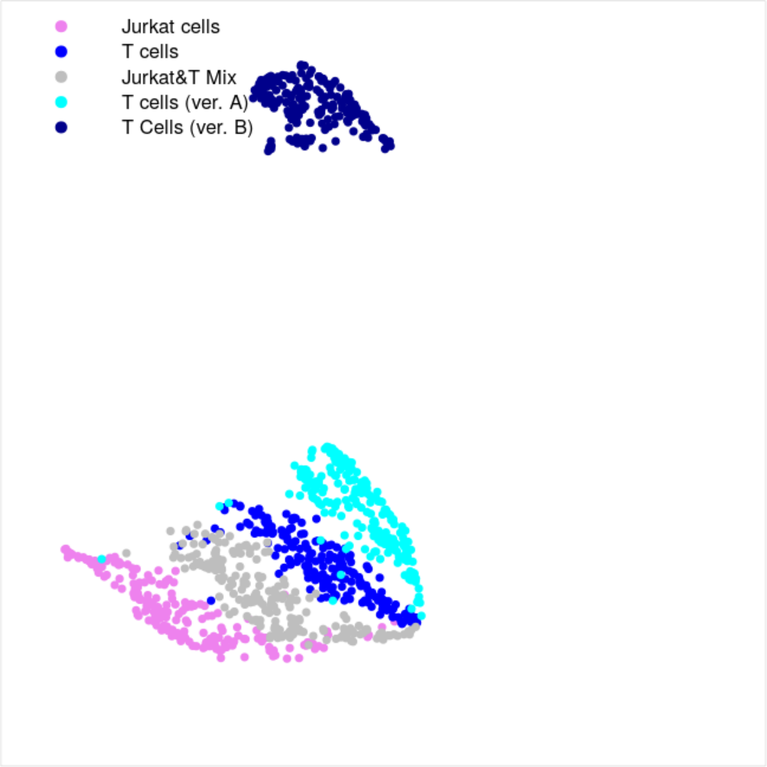
- 19536 genes
- 200 cells
- “new” T cells ver. B (many DE genes, large differences)

Simulation example 5

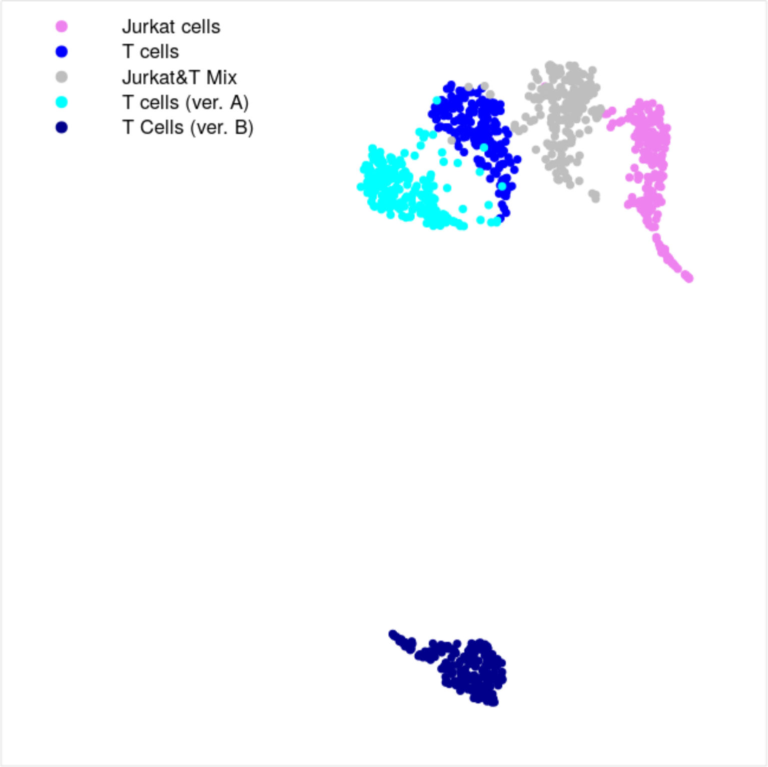
Gene expression level



Raw count



Normalized count


● Jurkat cells
● T cells
● Jurkat&T Mix
● T cells (ver. A)
● T Cells (ver. B)